Department of Health and Human Services
Public Health Service
Food and Drug Administration
Center for Drug Evaluation and Research
Office of Surveillance and Epidemiology Review (OSE)
Office of Pharmacovigilance and Epidemiology (OPE)

Epidemiology: Review of Extended-Release/Long-Acting (ER/LA) Opioid Analgesic PMR 3033-7 Final Study Report

Date:	November 13, 2019
Reviewer:	D. Tyler Coyle MD, MS *Division of Epidemiology II*
Team Leader:	Cynthia Kornegay PhD *Division of Epidemiology II*
Division Director:	Judy Staffa PhD, RPh *Associate Director for Public Health Initiatives* *Office of Surveillance and Epidemiology*
Drug Name:	ER/LA Opioid Analgesics
Subject	Review of ER/LA Opioid Analgesics PMR 3033-7 Final Study Report
Application Type/Number:	Multiple
Applicant/Sponsor:	Opioid PMR Consortium
OSE RCM #:	2016-367

TABLE OF CONTENTS

EXECUTIVE SUMMARY

The United States Food and Drug Administration's (FDA) Division of Anesthesia, Analgesia, and Addiction Products (DAAAP) asked the Division of Epidemiology II (DEPI) to review a report submitted by the Opioid Postmarketing Consortium (OPC) to fulfill a post-marketing requirement (PMR) to analyze abuse-related outcomes associated with the use of extended-release / long-acting (ER/LA) opioid products. FDA instructed the OPC to develop and validate an algorithm using coded medical terminologies to identify patients experiencing prescription opioid abuse or addiction (AA), among patients receiving an ER/LA opioid analgesic. The purpose of this document is to assess the epidemiologic methods, analysis, and results presented in the final study report for PMR 3033-7 (formerly Study 3B), entitled "An Observational Study to Develop Computable Algorithms for Identifying Opioid Abuse and Addiction Based on Administrative Claims Data."

The study's primary objective was to develop and evaluate a classification model based solely on medical claims data for identifying patients who have and have not experienced prescription opioid AA, as compared to the gold standard of manual chart review. This study was a retrospective observational cohort study using secondary data originating from medical claims records and electronic health records associated with the delivery of healthcare to patients in a variety of healthcare settings from 2006-2015. The primary study site was Kaiser Permanente Washington (KPW), and the secondary study sites were Kaiser Permanente Northwest (KPNW), Optum, and Tennessee Medicaid program, referred to as TennCare. The study population included individuals aged ≥18 years who received ≥60 days' supply of ER/LA opioid analgesics within any 90-day period during the study period.

Several developed algorithms tested different time periods of assessment; different combinations of claims codes; and use of natural language processing (NLP) extraction of data from electronic health records (EHR) to supplement the code-based algorithm for detecting AA using coded medical terminologies. Another algorithm was developed to predict the onset date of AA. The study operationalized candidate predictor variables as computable measures based on anticipated clinical relevance to generate risk scores comparing AA-positive to AA-negative samples. Sensitivity, specificity, negative predictive value, and positive predictive value of the algorithm were computed within the study site validation sample. Risk score cut-offs were modeled to maximize both sensitivity and positive predictive value for sample populations.

The algorithms demonstrated generally high levels of specificity across all study sites. However, **the algorithm had poor performance at discriminating between patients experiencing AA and those not experiencing AA in both the primary and secondary study sites.** The risk score cut-offs balancing sensitivity and positive predictive value for all algorithms developed resulted in unacceptably low positive predictive values. In many analyses, the algorithm performed little better than chance with respect to these metrics.

This study did not show that an automated algorithm based on medical claims data could determine which patient charts contained evidence of AA with high sensitivity and high PPV, nor that AA onset could be reliably determined through such an algorithm. **The findings of this study suggest that opioid abuse and addiction (AA) is a phenomenon that is not well-reflected in medical claims data.** Supplementing the algorithm with NLP- and EHR-generated data did not markedly improve performance. Compared to the gold standard of chart review, the developed claims-based algorithm was not able to identify the presence of opioid AA with sufficient accuracy to warrant further use. **The algorithm developed in this study should not be used further for any ER/LA PMR studies.**

Despite the algorithm's poor performance, adequate effort was made to develop an algorithm that successfully identified patients with AA using claims data. The postmarketing requirement should be considered fulfilled based on this submission, and the authors should strongly consider publishing these results so that other investigators can build on this work.

1 INTRODUCTION

The United States Food and Drug Administration's (FDA) Division of Anesthesia, Analgesia, and Addiction Products (DAAAP) asked the Division of Epidemiology II (DEPI) to review a report submitted by the Opioid Postmarketing Consortium (OPC) to fulfill a post-marketing requirement (PMR) to analyze abuse-related outcomes associated with use of extended-release / long-acting (ER/LA) opioid products. Specifically, the PMR required development and evaluation of potential code-based algorithms to identify patients experiencing opioid analgesic abuse or addiction among patients prescribed ER/LA products. The purpose of this document is to assess the epidemiologic methods, analysis, and results presented in the report.

2 REVIEW METHODS AND MATERIALS

2.1 DOCUMENT TO BE REVIEWED

On April 28, 2017, OPC submitted a final study report for PMR 3033-7 (formerly Study 3B) entitled "An Observational Study to Develop Computable Algorithms for Identifying Opioid Abuse and Addiction Based on Administrative Claims Data," dated April 18, 2017.

2.2 CRITERIA APPLIED TO REVIEW

The PMR letter to the OPC stated that study 3033-7 should be:
"An observational study to develop and validate an algorithm using coded medical terminologies to identify patients experiencing prescription opioid abuse or addiction, among patients receiving an ER/LA opioid analgesic."[1]

[1] Accessed on 5/24/2017 from https://www.fda.gov/downloads/Drugs/DrugSafety/InformationbyDrugClass/UCM484415.pdf.

The report was reviewed to assess this study’s ability to meet the goals defined in the PMR from an epidemiologic, methodologic, and analytic perspective.

3 REVIEW RESULTS

3.1 Study Objectives and Hypotheses

The primary objectives of this study were:

1) To create a high quality, manually-validated gold standard classification of opioid abuse/addiction (AA) for all patients in the study cohort using natural language processing (NLP) -assisted manual review of information extracted from the patients’ electronic healthcare record (EHR).
2) To develop and evaluate a classification model based solely on medical claims data for identifying patients who have and have not experienced prescription opioid AA, as compared to a high quality gold standard.

The secondary objectives of this study were:

1) To assess whether and to what extent the classification model referenced above could be used to ascertain the onset date of prescription opioid AA. This outcome depended in part on the availability of information within patient charts documenting the onset of abuse/addiction.

2) To develop, evaluate, and compare to the claims-data-only model developed for primary objective 2, to two alternative models designed to classify patients on the same outcome: a simple model based on a narrow set of diagnostic codes widely used in prior studies to identify problem opioid use, and a “best case” model using all available EHR data to identify patients with and without prescription opioid AA. The simple model was based entirely on a set of 14 International Classification of Disease (ICD)-9 diagnostic codes that have been used in prior studies to identify patients with problem prescription opioid use. The best case model included all data used in the model developed for primary objective 2, plus additional structured information from the EHR that is not available in claims data (e.g., laboratory study results) as well as information extracted from patients’ clinical encounter notes using NLP methods. Both the simple and “best case” models were developed and evaluated in this study.
3) To conduct a portability assessment to determine the performance characteristics of the model developed for primary objective 2 in three other settings, Kaiser Permanente Northwest, Optum, and patients receiving care through Tennessee Medicaid (TennCare) program.

The study’s hypotheses were:

Hypothesis 1: The investigators could determine with high reliability which patient charts had evidence of prescription opioid AA and which did not.

Hypothesis 2: Using medical claims data, the investigators could develop a fully automated algorithm that could determine, with high sensitivity and high positive predictive value, which patients had evidence of prescription opioid AA in their clinical charts and which patients did not.

Hypothesis 3: Using medical claims data the investigators could develop a fully automated algorithm that could determine, with high accuracy, the onset date of prescription opioid AA (defined as the date when evidence of this condition first appeared in a patient's clinical chart).

Hypothesis 4: Using medical claims data, EHR data, and information extracted from free-text clinical notes using automated NLP methods, the investigators could develop a fully automated algorithm that could determine, with high sensitivity and high positive predictive value, which patients had evidence of prescription opioid abuse/addiction in their clinical charts and which patients did not.

Hypothesis 4.A. The algorithm in Hypothesis 4 would perform significantly better than the algorithm in Hypothesis 2.

Reviewer Comment:

The stated primary objectives are consistent with the goals of the PMR. The exploration of portability to other data settings is appropriate and has the potential to meaningfully inform the ability of the algorithm to detect opioid AA in different claims environments.

While interesting, it is not clear that secondary objective 1 advances the PMR's goals.

The hypotheses are consistent to inform the goals of the PMR.

3.2 STUDY METHODS

3.2.1 Study Type

This study was a retrospective observational cohort study based on secondary use of data originating from medical claims records and EHR associated with the delivery of healthcare to patients in a variety of healthcare settings during the nine and one-half year period beginning January 1, 2006 and ending June 30, 2015.

Reviewer Comment:

This study type is appropriate to evaluate the defined objectives.

3.2.2 Data Sources

3.2.2.1 Primary Study Site

The primary study site was Kaiser Permanente Washington (KPW), a mixed-model healthcare system delivering outpatient primary and specialty care to over 475,000 patients in the state of Washington during the study period. KPW is an HMO-like system that uses Epic© EHR to document patient encounters. Inpatient care was provided by non-KPW providers reimbursed through contracted care plans or medical claims.

KPW patients, including those covered through the HMO plan, may receive care in <u>any</u> urgent or emergency room facility, and such encounters will not be documented in the KPW EHR. However, medical claims data from these encounters will be represented in the KPW enterprise-wide data warehouse (EDW). The KPW EDW is an enterprise-wide amalgam of structured data from all KPW data sources, including the EHR, medical claims, and other administrative data systems created and used for a variety of secondary purposes, including approved research. The Sentinel Common Data Model (SCDM) is a transformation of KPW EDW data that has been maintained and used to support a wide variety of research projects since 2004 and is currently compliant with the SCDM version 6 data model.

This study used EHR, EDW, and SCDM data in the computer-assisted manual chart review. SCDM data were used in developing the algorithms developed as primary and secondary objectives. A combination of SCDM and EHR data were used in developing the "best case" algorithm.

KPW contracts with external hospitals to provide inpatient care. It also contracts with external chemical dependency treatment facilities for chemical dependency treatment. This means that while KPW researchers do not have access to full-text encounter notes for inpatient and chemical dependency treatment, coded data related to such care – including diagnosis and procedure codes – are available through the KPW EDW. Notably, medications dispensed in connection with a patient's externally-contracted chemical dependency treatment are dispensed and documented through the KPW pharmacy.

Reviewer Comment:

Overall, KPW HMO represents a reasonable data system in which to conduct this study. It captures beneficiary encounters across multiple settings of care, as well as pharmacy data.

Patients enrolled in the Kaiser Permanente insurance system may represent a low-risk population. Between the higher socioeconomic status of the average beneficiary, Kaiser's institutional opioid use policy described below, and WA's unique state regulations for opioid prescribing, individuals in KPW may not be representative of the "average" person at risk for developing opioid AA due to baseline characteristics and regulated aspects of clinical care. Nevertheless, Kaiser represents a valuable data system and is helpful for informing the questions underpinning this PMR study.

Limitations of the data source include its limited geography to one state, and possible gaps in inpatient, emergency department, or substance treatment center encounters. However, the structure of KPW EDW captures all submitted claims, which may reduce the frequency of these missing data.

3.2.2.1.1 Population and Time Period

Patients included in this study:

a) Were ≥18 years of age by 1/1/2006, which was the start of the study period;

b) Received ≥60 days' supply of ER/LA opioid analgesics within any 90-day period; including transdermal or oral opioids but excluding buprenorphine, where the date of the first dispensing included in this medication episode was defined as the patient's study index date;

c) Had ≥6 months of continuous enrollment prior to the ER/LA medication-defined index date; *and*

d) Had ≥18 months of continuous enrollment following the ER/LA medication-defined index date *or* had at least 90 days of continuous enrollment after the index date and was known to expire between 90 days and 18 months following the index date.

Additionally, and only for the primary study site (KPW), patients were required to have at least eight calendar quarters during the study period (1/1/2006-6/30/2015) in which the KPW EHR contained clinical notes documenting some type of encounter. This criterion was imposed to increase the likelihood that sufficient information was present in each study patient's KPW electronic chart to render determinations regarding the presence or absence of AA based on manual chart abstraction.

Historical data show that 88% of KPW Integrated Group Practice (IGP; an HMO model of care) patients enrolled on a particular date will be continuously enrolled throughout the following year, and 77% will be continuously enrolled throughout the two-year period centered around that date (i.e., one year before and one year after the date of interest). Requiring at least 24 months of continuous enrollment thus maximized availability of complete patient data with a modest loss of generalizability to all KPW patients.

Patients were excluded from this study if they:
a) were residents of a nursing home at any time during the study period; or

b) were enrolled in a hospice care program at any time during the study period.

These exclusions were designed to prevent patients receiving opioids for palliative care from being included in the study because AA is not a primary concern in palliative care.

Reviewer Comment:

It is debatable whether individuals residing in a nursing home during the study period should have been wholly excluded. Many patients who undergo surgical procedures may spend time in a rehabilitation facility post-operatively during which they may be exposed to opioids. Depending on how the investigators defined a "nursing home," this may have excluded a relatively high-risk population: infirm individuals in the post-operative setting who are exposed to potent opioid analgesics.

With the above exception, the inclusion and exclusion criteria appear appropriate to identify and analyze the population of interest.

3.2.2.2 Samples and Sampling Procedure

Prescription opioid AA is a relatively rare condition among KPW patients. Therefore, to enhance statistical power, the analysis oversampled patients believed to be at elevated risk of opioid AA. There were two categories of risk specified:
a) Patients ≤35 years of age at study index date; and

b) Patients ever receiving one or more of the 304.*, 305.*, or 965.* diagnosis codes for opioid dependence, abuse/misuse, or poisoning (**Table 3.2.2.2.1**).

Because patients with these characteristics are a minority of the study population, the sample was weighted to over-represent these individuals.

Table 3.2.2.2.1. ICD-9 diagnosis codes used to define elevated risk of prescription opioid abuse/addiction in Study 3B.

ICD-9 code	ICD-9 description
304	Opioid Dependence, Unspecified
304.01	Opioid Dependence, Continuous
304.02	Opioid Dependence, Episodic
304.03	Opioid type dependence, in remission
304.7	Opioid/Other Dependence, Unspecified
304.71	Opioid/Other Dependence, Continuous
304.72	Opioid/Other Dependence, Episodic
305.5	Opioid Abuse, Unspecified
305.51	Opioid Abuse, Continuous
305.52	Opioid Abuse, Episodic
305.53	Opioid Abuse, in remission
965	Poisoning by opium (alkaloids), unspecified
965.02	Poisoning by methadone
965.09	Poisoning by other opiates/narcotics

Source: Final Report, Pages 17-18.

Reviewer Comment:

This ICD-9 code list seems appropriate to capture the outcomes of interest. The analysis does not provide details on the methods of weighting employed, which could be helpful for interpreting the final results.

3.2.2.3 Secondary Study Sites

This project included a portability study intended to generate information useful for optimizing the feasibility of implementing and calibrating the AA algorithm in three diverse study sites outside the primary site. The secondary study sites were:
1) Kaiser Permanente Northwest (KPNW);
2) Optum; and
3) Tennessee Medicaid program, referred to as TennCare.

Secondary study sites were selected because they represented diverse of health care settings, included populations cared for under fee-for-service commercial insurance arrangements (Optum), staff-model managed care (KPNW), and Medicaid (TennCare).

The primary study site provided detailed specifications to the secondary study sites regarding the selection of samples of up to 500 patients at each secondary study site. These samples were used to conduct analyses, specified by the primary site in collaboration with the secondary study sites, that informed the development of the AA algorithm in the primary site. Secondary study sites varied in their ability to sample patients' representative of their entire patient populations and in their ability to conduct manual chart reviews of complete patient charts. This impacted the characteristics of the resulting gold standard determinations regarding the presence or absence of AA rendered for sampled patients at each site. Portability analyses conducted at each secondary site were therefore adapted in a manner that was consistent with the available gold standard data. Calibration of the final AA algorithm based on site-specific data similarly varied across the three secondary sites based on available gold standard data.

Secondary sites are described below in detail.

Reviewer Comment:

It is unclear why secondary sites had difficulty performing manual chart reviews, and whether this difficulty was a function of inadequate funding or of the study timeline. It is also unclear why the secondary study sites had variable abilities to provide representative patient samples.

3.2.2.3.1 Kaiser Permanente Northwest (KPNW)

KPNW is an integrated health care delivery system providing integrated outpatient (primary and specialty), inpatient, and drug dependency treatment to approximately 540,000 individuals in Southwest Washington and Northwest Oregon. Salaried physicians working in health-plan-owned facilities provided the majority of patient care. Claims data capture any outside care. A common EMR system is used at all KPNW clinics, and comprehensive data from all patient encounters (e.g., demographics, diagnoses, procedures, and laboratory tests and results) are captured in clinical and administrative databases. Complete data on prescribed outpatient medications is also captured.

KPNW has an active Opioid Management Program in place, and is currently working to transition as many people to lower dose opioids (less than 120 morphine equivalent dose/day [MEQs]) as possible, and to convert people to long-acting opioids for chronic pain, either reducing or eliminating immediate-release opioids. The goals of these activities are to improve safety while also improving pain control. KPNW also has a pain clinic and members who are not responding to treatment in primary care are referred to that clinic. In addition, Oregon law states that prior to treating patients for "intractable pain" with controlled substances, clinicians are required to obtain signed documentation from the patient acknowledging the risks associated with opioid treatment. In compliance with the Oregon State requirement, KPNW has patients sign agreed Opioid Therapy Plans, and it does the same for members who are residents of Washington.

Washington State requires review and permission for all opioid prescriptions over 120 MEQs, thus KPNW's efforts to keep prescriptions under this limit are consistent with Washington's requirements. Nevertheless, a one-day snapshot (in October 2014) of musculoskeletal pain patients on stable doses of opioids for ≥90 days found 16% with daily prescriptions for >120 MEQs.

Reviewer Comment:

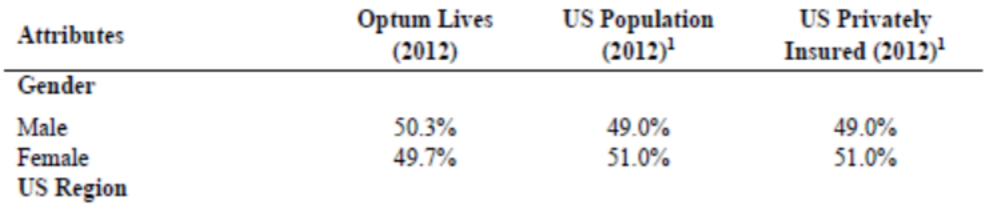
Between the higher socioeconomic status of the average beneficiary, Kaiser's institutional opioid use policy described above, and the unique state prescribing regulations seen in OR and WA, individuals in KPW and KPNW may not be representative of the "average" person at risk for developing opioid AA due to baseline characteristics and regulated aspects of clinical care. Nevertheless, Kaiser represents a valuable data system and is helpful for informing the questions underpinning this PMR study.

3.2.2.3.2 Optum

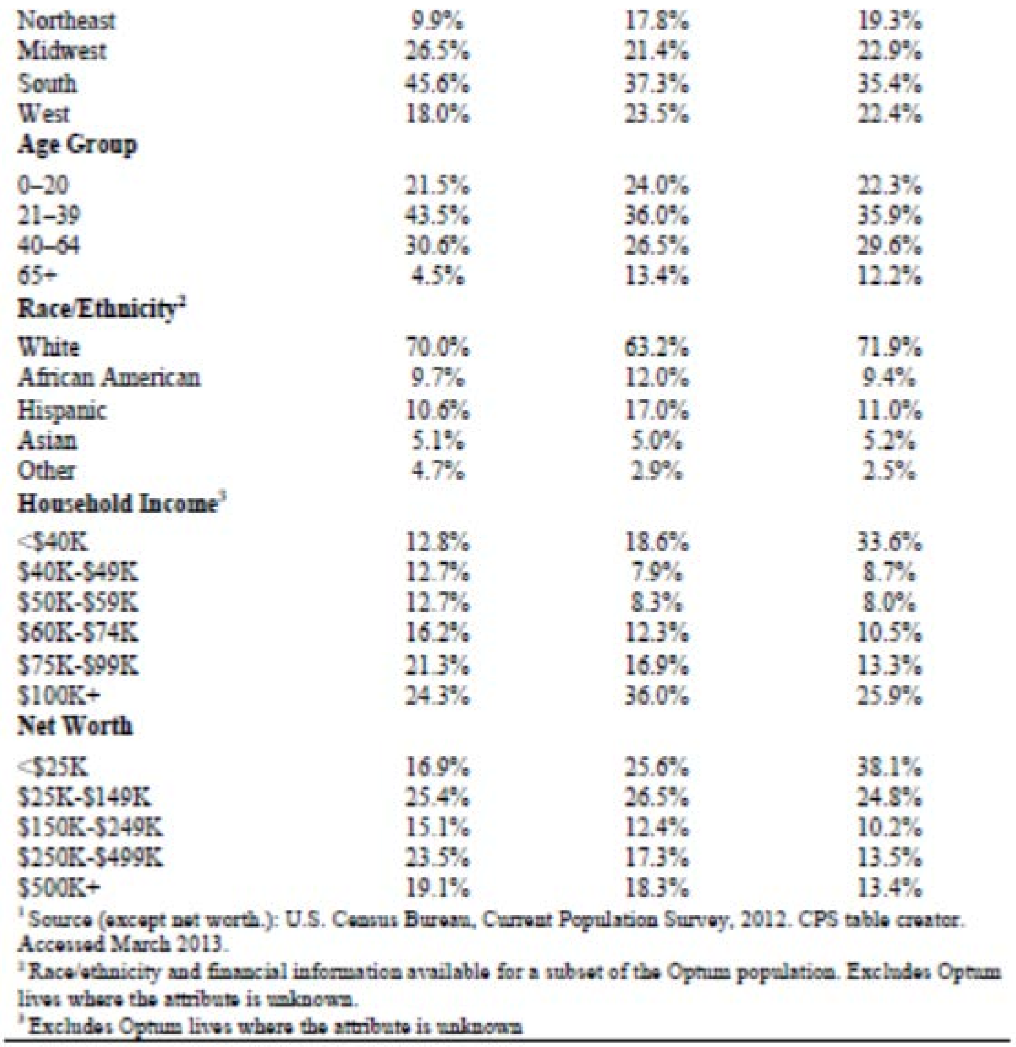
Optum has access to the Optum Research Database (ORD), which contains eligibility, pharmacy, and medical claims data from United Healthcare, supplemented by data derived from the EHR of a large U.S. commercial health plan affiliated with Optum (referred to as Humedica data). The ORD contains data relating to approximately 12.8 million individuals with both medical and pharmacy benefit coverage. The underlying population of insured persons is well-defined. The ORD contains protected health information and can be linked with appropriate approvals to external data sources, such as the U.S. National Death Index (NDI) or state vital statistics registries. In a subset of the ORD integrated with Humedica data, terms can be specified to be extracted from EHRs using NLP. Data extracted from EHRs do not include patients in the Optum database who are not insured directly through Optum's affiliate healthcare provider. ORD contains data on race/ethnicity and financial resource information for approximately 75-85% of the individuals included.

In this data, males and females were similarly represented, the Midwest and South census regions were overrepresented and the West was underrepresented. The Northeast had a substantially smaller representation in the data than other regions. Persons aged <65 years were distributed across ages roughly proportional to the U.S. as a whole, and persons aged >65 years are underrepresented. Household net worth indicates that the very poor are underrepresented.

Table 3.2.2.3.2.1. Demographic characteristics of persons with full medical and pharmacy claims data in the Optum Research Database in 2012 in comparison to the U.S. population and U.S. residents with non-governmental health insurance.



Attributes	Optum Lives (2012)	US Population (2012)[1]	US Privately Insured (2012)[1]
Gender			
Male	50.3%	49.0%	49.0%
Female	49.7%	51.0%	51.0%
US Region			



Northeast	9.9%	17.8%	19.3%
Midwest	26.5%	21.4%	22.9%
South	45.6%	37.3%	35.4%
West	18.0%	23.5%	22.4%
Age Group			
0–20	21.5%	24.0%	22.3%
21–39	43.5%	36.0%	35.9%
40–64	30.6%	26.5%	29.6%
65+	4.5%	13.4%	12.2%
Race/Ethnicity[2]			
White	70.0%	63.2%	71.9%
African American	9.7%	12.0%	9.4%
Hispanic	10.6%	17.0%	11.0%
Asian	5.1%	5.0%	5.2%
Other	4.7%	2.9%	2.5%
Household Income[3]			
<$40K	12.8%	18.6%	33.6%
$40K-$49K	12.7%	7.9%	8.7%
$50K-$59K	12.7%	8.3%	8.0%
$60K-$74K	16.2%	12.3%	10.5%
$75K-$99K	21.3%	16.9%	13.3%
$100K+	24.3%	36.0%	25.9%
Net Worth			
<$25K	16.9%	25.6%	38.1%
$25K-$149K	25.4%	26.5%	24.8%
$150K-$249K	15.1%	12.4%	10.2%
$250K-$499K	23.5%	17.3%	13.5%
$500K+	19.1%	18.3%	13.4%

[1] Source (except net worth.): U.S. Census Bureau, Current Population Survey, 2012. CPS table creator. Accessed March 2013.

[2] Race/ethnicity and financial information available for a subset of the Optum population. Excludes Optum lives where the attribute is unknown.

[3] Excludes Optum lives where the attribute is unknown

Source: Final Report, Pages 20-21.

Among the insurance contracts that lead to data being present in the ORD, almost all cover fee-for-service payments for health costs incurred by the covered individuals. These are commercial insurance policies obtained through small to medium-sized employers, and cover both the employee-contract-holder and that person's qualified dependents. The insurer seeks to modify physician and patient behavior through copay structures for drugs and services, by monitoring insurance claims for evidence of fraud or abuse and monitoring providers for wide deviations from community norms of cost, by sending health information to contract holders and by offering optional nurse counseling services. None of these practices are highly restrictive for drugs or services at the population level. Optum believes that the data are informative about the very broad segment of the U.S. population that could roughly be described as working, middle class and with the benefit of commercial health insurance.

Optum does not have an Opioid Management Program. However, it has introduced a variety of controls to promote good prescribing and utilization practices overall.

Reviewer Comment:

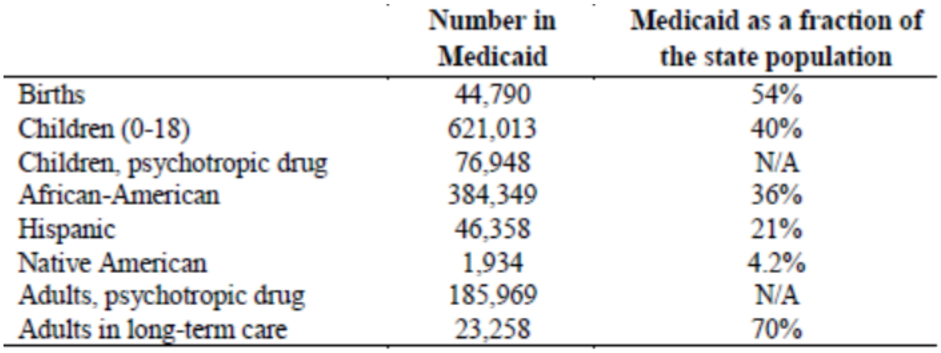
The report does not specify what prescribing and utilization controls have been implemented by Optum. Additionally, beyond gender and age distribution, it is not clear how similar Optum's patient population is compared to the overall U.S. population, or the privately-insured population. This makes it difficult to interpret these results due to concerns about generalizability.

3.2.2.3.3 TennCare

Tennessee Medicaid, TennCare, is a state-based, capitated, managed health-care program, covering 1.2 million Medicaid-eligible, uninsured, and uninsurable state residents. This Medicaid managed-care program maintains a computerized registry of all enrollees and records of patient-provider encounters and pharmacy benefits usage that allow the reconstruction of medication exposures and the identification of study outcomes. The following administrative data from TennCare have been computerized: an enrollment file, a pharmacy file that captures filled outpatient prescriptions, an inpatient file, and an outpatient file that includes encounter records for emergency department, hospital outpatient, and physician visits. Researchers from Vanderbilt University have access to TennCare data.

Medicaid is a joint federal-state program that finances medical care for four broad categories of low-income persons: parents and their dependent children, the disabled and blind, those aged 65 years and older. The 1.2 million enrollees in the Tennessee Medicaid program constitute 20% of the state's population, but account for approximately one-half of all births. Over a third of Medicaid beneficiaries are aged ≤18 years, and approximately a quarter are African-American or Hispanic. One in nine of all Tennessee Medicaid recipients in 2007 was an adult receiving psychotropic medication.

Table 3.2.2.3.3.1. Characteristics of the Tennessee Medicaid (TennCare) population in 2007.



	Number in Medicaid	Medicaid as a fraction of the state population
Births	44,790	54%
Children (0-18)	621,013	40%
Children, psychotropic drug	76,948	N/A
African-American	384,349	36%
Hispanic	46,358	21%
Native American	1,934	4.2%
Adults, psychotropic drug	185,969	N/A
Adults in long-term care	23,258	70%

Source: Final Report, Page 22.

For the purposes of this study, Tennessee Medicaid complemented other sites in its entirely low-income population, high proportion of African-Americans, geographic location, and high proportion of persons using psychotropic medications.

TennCare has implemented an Opioid Management Program in accordance with the State of Tennessee's Controlled Substance Monitoring Act of 2002. Accordingly, the Tennessee Department of Health established a database to monitor the dispensing of Schedule II, III, IV, and V controlled substances. Data collection in the Tennessee

Controlled Substance Monitoring Database (CSMD) began for all dispensers on December 1, 2006. The Prescription Safety Act of 2012 enhanced the monitoring capabilities of the database and added specific requirements.

Starting in 2013, all prescribers with DEA numbers who prescribe controlled substances and dispensers in practice providing direct care to patients in Tennessee for more than 15 calendar days per year have had to be registered in the CSMD. Pharmacies within the state of Tennessee are required to upload all schedule II-V prescriptions at least every seven days. All providers are required to review the recorded prescriptions in CSMD before initiating treatments with opioids or benzodiazepines. In addition, by State mandate, beginning in October 1 2013, no prescriptions for any opioids or benzodiazepines may be dispensed in quantities greater than a 30-day supply. In September 2014, the Tennessee Department of Health issued guidelines for outpatient management of chronic non-malignant pain to support clinicians' decisions for the treatment of patients with chronic pain.

In addition to the actions taken by the State, TennCare has progressively implemented specific actions to limit opioid exposure. During the last ten years, prior authorization has been implemented for the use of ER/LA opioid analgesic formulations. Furthermore, restrictions in the quantity of opioids dispensed have been implemented. Of note, all TennCare prescriptions have a maximum of 30 days of supply.

Reviewer Comment:
TennCare represents a relevant data environment in which to assess the algorithm among Medicaid beneficiaries. It should be noted that while dispensers were required to submit data to the CSMD starting in 2006, penalties for not reporting were enacted until 2013. Therefore, the system may not have captured all dispensed controlled substances in the state during the study period.

3.2.2.3.4 Inclusion Criteria

Inclusion criteria used to identify patients in the secondary study sites were the same as the criteria used in the primary study site, with the following exceptions.

a) Optum applied study eligibility criteria using United Healthcare claim data, and additionally required that a patient's ER/LA index date occur on the first day of at least a one-year period in which the patient's clinical encounters were documented by the Humedica EHR system. The overlap with Humedica data was designed to assure relevant information from clinical encounters for at least a year following the study index date be represented in the Optum profile.

b) TennCare limited its patient sample those receiving care at The Vanderbilt Clinic on the campus of the Vanderbilt University Medical Center and who met all of the following criteria:

    a. Had at least 70% of their TennCare visits at The Vanderbilt Clinic; and

    b. Had at least 2 visits to Vanderbilt within a 36-month period that began 12 months prior to study index date.

Both of the above adaptations were implemented to assure availability of information needed for the manual records review component of the portability study, which was used to establish gold standard determinations with respect to the presence or absence of AA for each patient in a site's sample. KPNW did not impose any additional inclusion criteria.

Reviewer Comment:

The restriction of the Medicaid sample to only patients receiving the majority of their care at a major university medical center may limit generalizability of the results: poor and disabled people who live near an academic teaching hospital may have a different AA risk profile than those who live in areas with access to fewer healthcare resources.

The other restrictions imposed appear to be reasonable for the purposes of ensuring comparable analyses across data systems.

3.2.2.3.5 Samples and Sampling Procedure

Simple random sampling was used to identify a set of patients to be included in the portability study sample at each secondary site. The planned sample size for each site was 500 patients. Investigators chose this number because it was the largest sample for which manual chart reviews could be completed within the established timeline.

The additional inclusion criteria applied at TennCare were required to assure minimally adequate availability of electronic patient charts to facilitate the manual abstraction component of the study. The investigators set these criteria based on their review of data summarizing the frequency and proportion of TennCare patient encounters occurring at The Vanderbilt Clinic. Though TennCare patients receive care at a large number of clinics throughout Tennessee, the investigators' Vanderbilt-based research team only had access to electronic patient charts at The Vanderbilt Clinic. Because of this limitation, the TennCare sample was far smaller than planned.

Reviewer Comment:

The unexpectedly small sample in TennCare – in addition to solely sampling from the Vanderbilt Clinic – further limits the interpretability of the results from this data stream.

3.2.2.3.6 Chart Abstraction and Gold Standard Creation

The study used manual chart abstraction to establish gold standard determinations for each patient in the study samples at the primary and secondary study sites. Manual chart abstraction was the preferred method because diagnostic coding represented in structured data is an unreliable indicator of AA. Chart abstraction procedures were specified in a detailed written protocol developed at the primary study site based on an iterative process involving review of 80 selected charts representing a variety of patient types and AA risk factors.

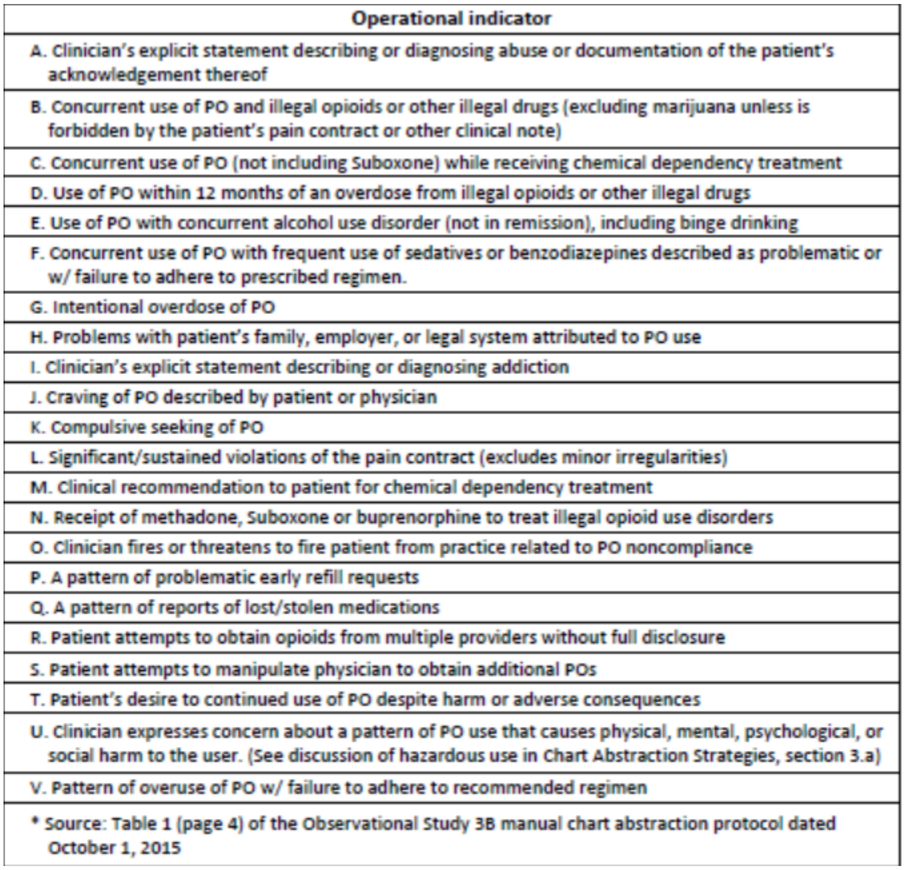
The written protocol described the process abstractors were trained in and instructed to follow in conducting abstraction of medical records for the study. It included a conceptual definition of prescription opioid AA that served as an overarching guide to abstractors' decision making. This conceptual definition was based on findings of the expert panel convened by the Analgesic, Anesthetic, and Addiction Clinical Trials, Translations, Innovations, Opportunities and Networks (ACTTION) partnership and

DSM-V criteria for prescription opioid use disorder. It defined prescription opioid abuse as the use of these medications for nontherapeutic purposes to obtain psychotropic (e.g., euphoric, sedative, or anxiolytic) effects, that contradicts medical advice, that causes physical, mental, psychological or social harm to the user, that is illegal, or that that constitutes sustained hazardous use. Hazardous use was defined as a pattern of substance use that increases the risk of harmful consequences for the user, including physical harms, mental harms, and social consequences. The protocol defined prescription opioid addiction (or prescription opioid use disorder) as the compulsive use of prescription opioids that occurs despite personal harm or negative consequences and may involve impaired control over use of and/or craving for the substance. References to relevant scientific and clinical literature for these definitions are included in the chart abstraction protocol, available as a separate document title "Manual Chart Abstraction Protocol for Observational Study 3B" dated November 30, 2015.

The chart abstraction protocol also describes the strategies and tactics abstractors use when abstracting a chart, a series of specific steps to be followed when searching a patient’s chart for relevant evidence, and procedures for recording required information in a study tracking database. For each chart reviewed abstractors made a binary determination regarding evidence of AA, recording a “yes” if evidence was present and “no” otherwise.

Abstractors were instructed to make these binary determinations based on the totality of evidence documented in the chart. In doing so, they considered a list of 22 operational indicators of prescription opioid abuse/addiction, listed in Table 3.2.2.3.5.1.

Table 3.2.2.3.5.1. Twenty-two operational indicators of prescription opioid (PO) abuse and/or addiction (AA) used as described in the Study 3B manual chart abstraction protocol.*



Operational indicator
A. Clinician's explicit statement describing or diagnosing abuse or documentation of the patient's acknowledgement thereof
B. Concurrent use of PO and illegal opioids or other illegal drugs (excluding marijuana unless is forbidden by the patient's pain contract or other clinical note)
C. Concurrent use of PO (not including Suboxone) while receiving chemical dependency treatment
D. Use of PO within 12 months of an overdose from illegal opioids or other illegal drugs
E. Use of PO with concurrent alcohol use disorder (not in remission), including binge drinking
F. Concurrent use of PO with frequent use of sedatives or benzodiazepines described as problematic or w/ failure to adhere to prescribed regimen.
G. Intentional overdose of PO
H. Problems with patient's family, employer, or legal system attributed to PO use
I. Clinician's explicit statement describing or diagnosing addiction
J. Craving of PO described by patient or physician
K. Compulsive seeking of PO
L. Significant/sustained violations of the pain contract (excludes minor irregularities)
M. Clinical recommendation to patient for chemical dependency treatment
N. Receipt of methadone, Suboxone or buprenorphine to treat illegal opioid use disorders
O. Clinician fires or threatens to fire patient from practice related to PO noncompliance
P. A pattern of problematic early refill requests
Q. A pattern of reports of lost/stolen medications
R. Patient attempts to obtain opioids from multiple providers without full disclosure
S. Patient attempts to manipulate physician to obtain additional POs
T. Patient's desire to continued use of PO despite harm or adverse consequences
U. Clinician expresses concern about a pattern of PO use that causes physical, mental, psychological, or social harm to the user. (See discussion of hazardous use in Chart Abstraction Strategies, section 3.a)
V. Pattern of overuse of PO w/ failure to adhere to recommended regimen
* Source: Table 1 (page 4) of the Observational Study 3B manual chart abstraction protocol dated October 1, 2015

Source: Final Report, Page 25.

Each operational indicator represented a type of evidence that may be relevant in determining whether a chart contained evidence of prescription opioid AA. These operational indicators were not used as a simple checklist for determining whether evidence of AA was present in a chart. Rather, they provided a rubric for abstractors to consider when judging the totality of evidence documented in a particular patient's chart. Strength of evidence represented by a given operational indicator often varied considerably from one chart to the next. For example, a series of five early refill requests during a short period of time when a patient's primary care physician is expressing concern about the patient's apparent inability to comply with their mutually agreed upon opioid care plan is much stronger evidence of a pattern of problematic early refill requests than a pattern of two early refill requests six months apart with no expressions of clinical concern in a patient's encounter notes. Abstractors were instructed to consider the strength of evidence in making their determinations. They were also instructed to carefully weigh potentially conflicting evidence in a patient's chart regarding operational indicators of abuse/addiction. For example, an Emergency Department (ED) doctor unfamiliar with a patient's ongoing primary care may record in an ED encounter note that the patient's request for opioids was evidence of problematic early refill requests (operational indicator P in Table 5.4), while the patient's primary care team, who had regular and timely encounters with the patient, considered the patient to be adherent to a mutually agreed upon management plan and considered the patient's ED encounter to be an isolated and reasonable occurrence. In such situations abstractors may have determined that a chart contained evidence of an operational indicator (e.g., the ED doctor's note about early refills) but conclude based on the totality of evidence in the chart that the patient was not experiencing AA. The chart abstraction protocol instructed abstractors that weak or incomplete evidence was not sufficient to support a determination that AA was present.

While operational indicators of AA provided a useful rubric for abstracting patient charts, and abstractors were instructed to record which operational indicators supported their determinations, abstractors were not asked to identify and record all operational indicators that may be present in a patient's chart. This is because charts of long-term opioid recipients are often voluminous. The median number of words per patient chart during the study period at KPW was 88,616, and the inter-quartile range for word count was 50,941-143,921. Attempting to exhaustively identify all operational indicators in such charts was unnecessary, was not feasible within the study timeline, and would have been of questionable value, considering that the only relevant information needed from the chart abstraction to achieve the study's primary objective of developing the classification algorithm was whether evidence of AA was present and if so when that evidence first appeared in the chart. In addition to recording for each chart the binary outcome regarding AA, abstractors also rated their self-assessed confidence level regarding each of their determinations. They did this using an ordinal scale of high confidence, medium confidence, or low confidence. All charts reviewed by their first assigned abstractor with medium or low confidence – regardless of the determination with respect to presence or absence of AA – was received a second, independent, blinded review by another abstractor who was unaware they were re-reviewing a previously reviewed chart. The process for resolving any discordances in the review process involved additional reviews, including as needed review by an adjudication committee.

The written chart abstraction protocol was adapted for use at the three secondary sites involved in the portability study. Adaptations were made to accommodate the protocol to site specific differences in locally available EHR data.

3.2.2.3.6.1 **Primary Study Site**

Developing and evaluating automated algorithms requires high-quality information with which to develop and validate those algorithms. A foundational component of the study is the creation of a high-quality reference standard (also referred to as a gold standard) for a large number of patient charts based on a systematic and thorough manual abstraction of patients' complete electronic medical records.

Details of the chart abstraction method used at the primary study site are provided in a written chart abstraction protocol available in Appendix 1. That protocol describes the process for conducting the abstraction of medical records for this study. It begins with a definition of prescription opioid AA used to guide abstractors' decision making. It also describes the strategies and tactics abstractors use when abstracting a chart, a series of steps to be followed when searching a patient's chart for relevant evidence, procedures for recording required information in a study tracking database. The process used to evaluate inter-rater reliability (IRR) is described in the statistical analysis plan (SAP, available as a separate document titled “Statistical Analysis Plan for Observational Study 3B” updated October 1, 2015).

3.2.2.3.6.2 Secondary Study Sites

In consultation with the primary study site, the secondary study sites used the protocol developed by the primary study site to create gold standard determinations for their respective patient samples, with the following adaptations.

a) None of the secondary sites used the NLP-assisted chart abstraction tools used at the primary study site because the expected costs of implementing these tools at each secondary site were believed to outweigh the expected benefits in terms of increased chart abstraction efficiency.

b) KPNW included inpatient and outpatient addiction treatment encounter notes in its review because such services are provided to KPNW patients through KPNW facilities and the resulting encounter notes are available in the KPNW EHR.

c) Optum conducted its entire review using a resource referred to as the patient profile. Optum patient profiles consist of an assembly of temporally-ordered structured claims data and structured data generated by an NLP system developed to extract and organize concepts from free-text into semi-structured fields. NLP items are derived from text entries that correspond primarily to terms in two large dictionaries, SNOMED and Med-DRA (Medical Dictionary for Regulatory Activities; see http://www.meddra.org/). Each NLP item consists of a concept – e.g., “nausea” – together with attributes derived from the immediate sentence context and from the location of the observation in the medical record. Additional features capture the location in the record (e.g., the “Subjective” section of a clinical note) and important contextual mentions such as family history or denial, should they be present.

d) Optum used three categories when rendering determinations regarding AA for each patient: 1) present, 2) possibly present, and 3) absent.

Because the Optum patient profile does not include actual, verbatim text from patient chart notes, and because of the additional uncertainty regarding evidence of AA (reflected in the use of a “possibly present” category), investigators considered the gold standard data for Optum patients to be most useful as a secondary estimate of AA prevalence rather than an unbiased measure of each patient’s true AA status.

Reviewer Comment:

The chart abstraction and gold standard creation procedures appear to be appropriate and sufficiently harmonized across the sites to support the study required by the PMR.

3.2.2.3.7 Learning and Validation Samples in the Primary Study Site

Investigators used stratified random sampling to divide the 2,000 study subjects into a 60% training sample (N=1,400) and 40% validation sample (N=600). Sampling was stratified on categories of the two binary risk indicators (age ≤35 years versus older, and presence of a diagnosis code for opioid dependence, abuse, or poisoning versus never) to assure balance on these characteristics in the samples. The training sample was used throughout all phases of algorithm training. The validation sample was reserved for validation of the final algorithms.

3.2.2.3.8 Algorithm Development and Validation in the Primary Study Site

The investigators developed two types of algorithm using data at the primary site. The first type consisted of classification algorithms to identify patients whose medical records contained evidence of AA. The investigators developed and evaluated five versions of these classification algorithms:

a) Full-period classification algorithm: This algorithm used predictor variables derived from medical claims data (only) available during the entire 9.5-year study period to determine whether each patient's record contained evidence of AA (referred to as AA positive) or not having (referred to as AA negative) evidence of prescription opioid AA.

b) 36-month classification algorithm: This is identical to the full-period algorithm except that the predictors were limited to claims data during a 36-month period beginning 12 months before each patient's study index date and ending 24 months after the index date.

c) Full-period ICD-9 classification algorithm: This is identical to the full-period algorithm except that predictors were restricted to a set of 15 ICD-9 codes for opioid-related dependence, abuse, and poisoning. Such algorithm have been widely used and reported in the literature.

d) 36-month ICD-9 classification algorithm: This is identical to the full-period ICD-9 algorithm except that the predictors were limited to the 15 ICD-9 codes during a 36-month period beginning 12 months before each patient's study index date and ending 24 months after the index date

e) 36-month EHR-enhanced classification algorithm: This is identical to the full-period algorithm except that the predictors based on claims data were supplemented with data only available in EHR systems and data extracted from patients' clinical notes via NLP.

The second type of algorithm was designed to predict the onset date of AA among patients who were not known to have experienced AA prior to their study index date. There was one version of this algorithm:

f) AA onset prediction algorithm: This algorithm used a time-to-event model to predict incident AA during an approximately two year period following eligible patients' study index date.

The methods used to develop and evaluate each of the above six algorithms at the primary site are described below.

Reviewer Comment:

The approaches to algorithm development appear to be appropriate and sufficiently varied to meaningfully inform the question underpinning the PMR.

It is not clear that the AA onset prediction algorithm furthers the goals of the PMR.

3.2.2.3.9 AA Classification Algorithms

The overall strategy for developing classification algorithms was to generate a large number of candidate predictor variables known or believed to be associated with the primary outcome (AA positive or AA negative), and then apply statistical techniques to select the best subset of those predictors as the basis of a model that generated a risk score, ranging from 0 to 1, indicating each patient's likelihood of being AA positive. The investigator then selected risk score cut points to dichotomize patients with respect to the outcome, i.e., AA positive versus AA negative. This section describes the process followed to develop and evaluate the full-period AA classification algorithm. Below, variations to the process are described to create the remaining four classification algorithms.

The full-period AA classification algorithm was developed as follows.

a. Candidate predictor variables that were known or believed to be associated with AA were specified based on 1) findings reported in the literature, 2) expertise of clinicians with extensive experience treating patients with chronic pain, 3) insights gained from manual abstraction process used to create the gold standard, and 4) prior experience developing algorithms for identifying patients with problem opioid use (e.g., morphine equivalent dose of opioid medications dispensed during a specified period of time).

b. Each candidate predictor variable was operationalized as a computable measure using training data. If there were several reasonable alternative ways to operationalize a candidate predictor, it was operationalized in ways (e.g., calculating morphine equivalent dose over one, two, or three months).

c. The relationship between each operationalized predictor and the gold-standard primary outcome (AA positive versus AA negative) was examined in training data to assess how well the predictor distinguished between AA positives and AA negatives. Investigators used this information to produce dichotomized versions of most continuous predictor variables, as continuous versions would often serve as proxies for length of enrollment. As length of enrollment is greatly variable between insurance settings, investigators did not believe that any findings based on it would be generalizable beyond managed care settings.

d. Investigators used the following ratio to identify operationalized versions of candidate predictor variables to include in the variable selection process:

% of AA **positives** where the predictor variable is positive ÷
% of AA **negatives** where the predictor variable is positive

Generally, if this ratio was ≥1.5 (i.e., at least 50% more AA positives were flagged by the predictor than AA negatives were flagged by the predictor), and the absolute number of AA positive patients flagged by the predictor was ≥10, the candidate predictor was included in the variable selection process. The study also included candidate predictor variables with high face validity, and predictor variables that operationalized age-group interactions with other candidate predictors. The study did not include among the set of candidate predictors measures investigators hypothesized would be related to AA status, but which turned out to exhibit no association in exploratory bivariate analyses. An example of candidate predictors not included were measures related to surgeries for specific anatomical locations (e.g., back surgery or neck surgery). After exploring the data for such measures in the training set, investigators found no differences between AA positive and AA negatives on these measures, and therefore did not include them in the set of potential predictor variables.

e. As both the training and validation data were oversampled for younger, high-risk patients, investigators developed inverse probability weights to reweight the analytic datasets back to the general population. For each of the four sampling groups (Risk status, age status), the inverse probability weight was (# general population)/(# analytic population). These weighting estimates are used in the LASSO modeling, as well as for evaluating model performance in both the training and validation samples.

Reviewer Comment:

LASSO stands for least absolute shrinkage and selection operator. It is a regression analysis method that performs variable selection to enhance accuracy of the model it produces. This may be a reasonable modeling approach to algorithm development. DEPI defers to our colleagues from the FDA's Office of Biostatistics regarding the appropriateness of modeling in this study.

f. The adaptive LASSO modeling function used was lqa() from the LQA R package. Prior to running the model, investigators needed to remove variables that were perfectly collinear with other combinations of variables in the dataset. In order to do this, investigators iterated through the dataset, checking to see if the addition of each subsequent variable increased the rank of the predictor matrix. (If the rank didn't increase, that means the variable can be exactly replicated by a combination of other variables, and thus it adds no information to the predictor matrix).

g. The adaptive LASSO requires coefficient weights to influence the speed at which its beta estimates go to zero due to the regularization effect of the penalty parameter lambda. LASSO models work by penalizing their goodness-of-fit metrics depending on how far their coefficient estimates are from zero. The adaptive portion makes it so that coefficients with high baseline values have their penalty lessened relative to those that have small effects. To get these values, investigators used ridge regression, as standard logistic regression would fail in the setting of 1,400 observations and around 1,000 predictors. The adaptive LASSO coefficient weights used are the inverse of the absolute value of the coefficients obtained from ridge regression.

h. The adaptive LASSO has two parameters that need to be specified to fit a model. The first is gamma – an exponent applied to the coefficient weights. Secondly, the lambda parameter is required, to influence just how much penalization is applied to the estimated coefficients – this is what makes LASSO models shrink coefficient values to zero, producing parsimonious models. To estimate these values, investigators used eight-fold cross validation on the training data, performing a grid search over values of both gamma and lambda. Investigators chose eight-fold due to the presence of indicators with rare events; it is likely that larger numbers of folds would produce cross-validation models with predictors that correspond to zero events. The metric for evaluating the fit given lambda and gamma was a sum of squares: sumi((yi-ŷi)2), where ŷ comes from the left-out portion of the cross-validation sample.

i. With estimates for both lambda and gamma, investigators were then able to fit adaptive LASSO models on our entire training dataset. These are the final models from this study.

Reviewer Comment:
DEPI defers to the assessment from the FDA's Office of Biostatistics on the above modeling approaches.

j. To assess the algorithm's overall performance investigators dichotomized patients based on selected cut points of the algorithm's risk score. **Patients were classified as AA positive if they had risk scores ≥ the cut point, and AA negative otherwise.** Investigators did this for a limited number of cut points chosen to optimize alternative performance characteristics: 1) desirable sensitivity, 2) desirable specificity, 3) desirable PPV, or 4) a balance between sensitivity and PPV.

During the algorithm development phase, investigators selected cut points based on training data. To evaluate the final algorithm, investigators selected cut points based on training data, and reported performance characteristics based on validation data. Key metrics used to evaluate the classification model were:

i. **Sensitivity** (also referred to as the recall rate or true positive rate) was defined as the ratio: true positives / (true positives + false negatives). A priori, investigators acknowledged trade-offs between sensitivity and specificity. <u>Ideally, investigators wanted a classification algorithm that achieved sensitivity ≥0.90 and specificity ≥0.90.</u>

ii. **Specificity** (also referred to as the true negative rate) was defined as the ratio: true negatives / (false positives + true negatives). As noted above, the trade-off between sensitivity and specificity is important to evaluate in the use of this algorithm.

iii. **Positive predictive value** (PPV, also referred to as precision) was defined as the ratio: true positives / (true positives + false positives). Use scenarios where false positives are expensive or otherwise unacceptable require high PPV.

iv. **Negative predictive value** (NPV) was defined as the ratio: true negatives / (true negatives + false negatives). Use scenarios where false negatives are unacceptable require high NPV.

k. To evaluate the final AA algorithm investigators reported its performance characteristics in validation data using risk score cut points selected based on training data that that the following performance characteristics:

i. Sensitivity (at levels considered excellent, good, or acceptable)

ii. Specificity (at levels considered excellent, good, or acceptable)
iii. PPV (at levels considered excellent, good, or acceptable)
iv. Balanced Sensitivity and PPV

l. To graphically assess the performance of the final AA algorithm, investigators produced receiver operating characteristic (ROC) curves showing the sensitivity-specificity tradeoff in both the training as well as the validation data.

Reviewer Comment: For a given population's ROC curve, sliding the risk score cut-off point along the curve changes test performance metrics: sensitivity, specificity, PPV, and NPV. It appears the investigators have created an analytic model that slides the risk score cut-off point along the ROC, and calculates the cut-off point that simultaneously maximizes/optimizes PPV and sensitivity for that population. This approach adds value to the algorithm by not simply identifying the most obvious cases of AA (high PPV) at the expense of missing cases that are less obvious (low sensitivity). However, the investigators do not describe this approach in detail, which is a notable deficiency.

The cutoffs for sensitivity, specificity, and PPV described in the results tables and referenced in subsection K above are generally acceptable.

3.2.2.3.10 Algorithm to Predict AA Onset (time-to-event)

To determine the feasibility of using a claims-based algorithm to identify incident AA, investigators developed and evaluated a time-to-event model as follows. First, investigators restricted each patient's observation period to a 720-day period (approximately two years) following their study index date. This period matched the duration of follow up used in the 36-month version of the AA classification algorithm (which included 12 months prior to and 24 months following a patient's study index date).

Patients included in the time-to-event analysis were those that met the inclusion criteria, and additionally had no evidence of AA prior to their study index date according to the manually abstracted gold standard. For purposes of the time-to-event model, such patients were considered AA positive if their AA onset date according to the gold standard occurred during the 720-day period following their study index date and AA negative otherwise. AA negatives included patients with AA onset dates >720 days after their study index date.

The investigators' goal was to predict AA onset within one of eight 90-day windows of time during the 720-day follow up period. The study therefore divided the 720-day follow up period into eight successive periods of 90-days each and mapped each AA positive patient's onset date to one of these eight periods. If the chart abstraction yielded only a year of onset (which was allowed according to the chart abstraction protocol if the abstractor determined evidence in the chart was not sufficient to establish a month and year of onset), investigators assigned the chart a randomly selected month.

Predictor variables for the time-to-event model were calculated for each patient (AA positives and AA negatives) separately for each of the eight 90-day follow up periods. Due to the smaller sample size that resulted from dropping individuals with AA prior to their index date, investigators made two concessions to prevent overfitting. First, instead of splitting the data into training and validation samples, investigators developed the model on the full body of data. This enabled use of all possible events in the model development, and production of stable parameter estimates. Secondly, rather than operationalizing quarter-specific variables for all the potential predictors studied in the main AA algorithm, investigators limited analysis to only looking at those predictors kept by the final algorithm.

Predictors were operationalized for each patient and for each period as follows:

Binary predictors

a) Binary flag indicating whether the predictor is TRUE or FALSE in the current period;

b) Binary flag indicating whether the predictor has ever been TRUE in the current period or prior periods (i.e., once this flag is set to true it remains true for all remaining periods)

Interval-level predictors

c) Value of the predictor in the current period;

d) Cumulative sum of the values of the predictor for all periods up to and including the current period

The time-to-event model was fit on a dataset with one row per patient-quarter, excluding quarters that occurred after censoring or incidence of AA. Again, adaptive lasso was used to estimate the likelihood that a patient would experience incident AA in that quarter, using ridge regression to obtain coefficient weights. With this model, person-quarter predictions for AA incidence were made on the full dataset, without censoring at the time of true AA incidence.

Investigators then calculated cumulative probabilities of predicted AA incidence per quarter, and identified a threshold at which the predicted prevalence of AA in the dataset matched the observed prevalence. The predicted quarter of AA incidence was the time at which the person’s cumulative probability exceeded this threshold. If the cumulative predicted probability never exceeded the threshold, the model predicted that the person would not have incident AA within the 720 days after the index date.

Investigators evaluated the time-to-event model by cross-tabulating and comparing patients’ predicted AA onset quarters (1 through 8 or never) against their gold standard onset periods (1 through 8 or never), and calculating the following quantities:

f) Among all patients, percentage agreement on the exact period of onset (or never);

g) Among all patients, percentage agreement within a window of time including the prior, current, and following period;

h) Among patients predicted to be AA positive during the follow up period (i.e., AA incident), percentage agreement on the exact period of onset;

i) Among patients predicted to be AA positive during the follow up period (i.e., AA incident), percentage agreement within a window of time including the prior, current, and following period.

Reviewer Comments:
It is not clear to DEPI that the AA onset algorithm is relevant for informing the PMR.

3.2.2.3.11 Portability Assessment in the Secondary Study Sites

Portability assessment activities were designed to gain knowledge about the data and patients in each of the secondary study requests so that this knowledge could be used, where appropriate, to influence development of the AA algorithm, which was based primarily on analyses conducted on KPW data, as described above. Investigators implemented these portability assessments with each secondary study site through a series of 13 “data requests” issued between December 2, 2015 and February 21, 2017. Each data request consisted of a set of written instructions for analyses to be conducted at each secondary study site, usually accompanied by SAS programming code that could be adapted for local use. Data requests were distributed to the study sites during the course of the study. Data requests addressed questions ranging from characterizations of patient demographics, the distribution of patients by study eligibility criteria, distribution of opioid medication utilization, distribution of selected diagnosis codes or categories of diagnosis codes, distribution of exposure to other types of medications (such as benzodiazepines), findings of the local chart abstraction for presence/absence of AA, and the results of implementing preliminary and final versions of the AA algorithm. Sites responded to each data request by transmitting summary tables and/or de-identified data to the primary study site, consistent with their respective data use agreements.

The portability SAP is included as Appendix 2 in the final report. Investigators chose to evaluate the AA model that used 36 months of data at each of the three external sites. A limited recalibration of parameters was carried out at KPNW, re-estimating the intercept on 500 individuals with chart-abstracted AA outcomes, as well as a parameter for adjusting all other model coefficients. At TennCare, only a new intercept was estimated, while no recalibration was carried out on Optum data. Model performance at the sites was summarized in tables showing sensitivity, specificity, PPV, NPV and predicted prevalence, as well by graphical displays of ROC curves.

3.2.2.3.12 Adaptations and Extensions

During the course of this project several unanticipated opportunities and challenges became apparent. In response, the Study 3B team, in consultation with Study 3A, Study 1B, and OPC/OSW members, adapted the originally planned work to best address these opportunities and challenges. This section describes each challenge or opportunity and the consequent adaptations or extensions implemented.

- **Adoption of an existing common data model**

Investigators originally planned to re-engineer data available from the KPW EHR system to simulate medical claims data. However, the 3A, 3B and 1B study teams decided to instead commit to using the SCDM as the data model for this collection of studies, so the study implemented and used data in the SCDM format instead.

- **Algorithm portability studies**

A portability study to assess the feasibility of implementing, in three secondary study sites, the AA algorithm developed in the primary study site was originally planned to be part of PMR Study 1B. However, Study 3A, 3B, and 1B teams decided, in consultation with the OPC/OSW, that greater value would be achieved by incorporating the portability studies for the 3B AA algorithm into Study 3B, thereby allowing information learned during the portability study to be incorporated into the 3B AA algorithm during its development phase. Investigators therefore added the portability study pertaining to Study 3B's AA algorithm (its primary objective) into Study 3B, and describe the results of that study here.

- **Optum gold standard**

Actual full-text encounter notes from the clinical charts of patients in the Optum sample were not available for manual review during chart abstraction to determine gold standard classifications regarding AA. Instead, patient "profiles," consisting of collections of chronologically-organized structured data codes and codes for terms from the SNOMED and Med-DRA medical dictionary identified in patient chart notes using NLP were used for records review at Optum. Given the complex nature of AA and the subtlety and indirectness with which clinicians often describe AA in patient charts, investigators considered the Optum records review results to be suitable for some but not all aspects of the AA algorithm development work. Investigators did not, for example, use the Optum records review results to re-calibrate the AA algorithm. Nevertheless, there was value in the Optum data that helped inform development of the AA algorithm.

- **TennCare sample**

It was not feasible for the Vanderbilt University research team to conduct chart reviews of TennCare patients receiving care in any clinic other than the one associated with Vanderbilt University. Further, a very limited number of TennCare patients received a substantial portion of their outpatient care at the Vanderbilt University Clinic. A decision was made to review the charts of 67 TennCare patients who received at least 70% of their care at the Vanderbilt Clinic. These 67 patients became the TennCare portability sample.

Because of its relatively small sample size (N=67 patients) and questions about the generalizability of patients receiving care at the Vanderbilt University clinic to the larger TennCare population, investigators limited use of the TennCare portability sample to a partial re-calibration of the final AA algorithm developed using KPW data.

Reviewer Comment:

The small sample size and generalizability issues surrounding the TennCare sample are important factors to consider when interpreting the results.

The lack of access to Optum medical data could also be problematic, as it is not known how well the NLP processing captures the terms and concepts of interest.

3.3 STUDY RESULTS

3.3.1 Primary Study Site

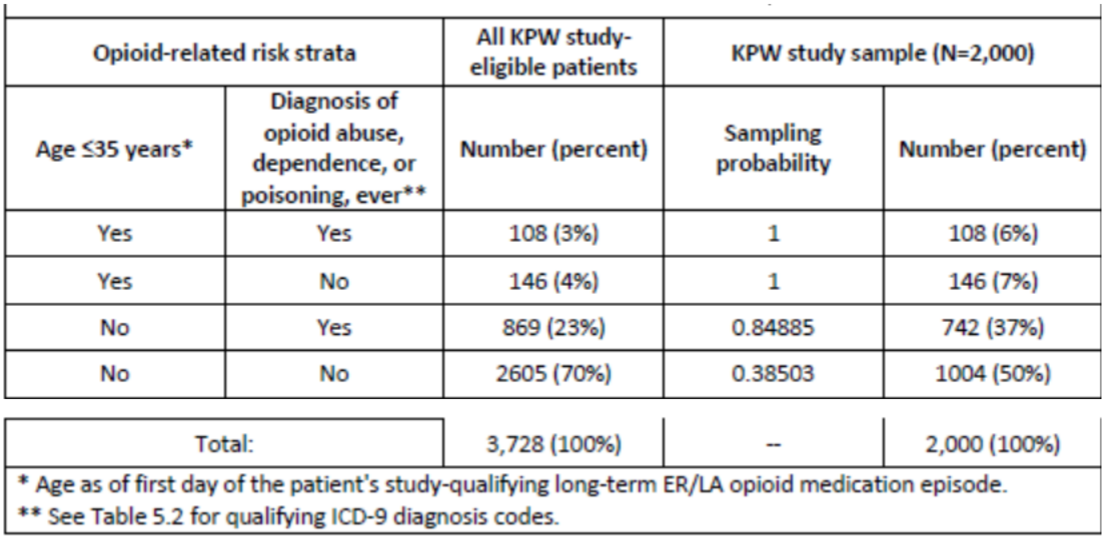
3.3.1.1 Sample Selection

A total of 3,728 KPW patients met all study inclusion and exclusion criteria. These patients all received care through the IGP (“HMO” model of care) in western Washington. This assured that their outpatient primary and specialty care would be documented in the KPW EHR. The median number of months in study-qualifying continuous enrollment periods (which required overlap with a study-qualifying episode of long-term ER/LA use) was 101 months (interquartile range [IQR]: 58-121, minimum 24, maximum 141). The majority of these patients had a single qualifying continuous enrollment period (median 1, IQR: 1-1, minimum 1, maximum 3). The median age was 52 years (IQR: 44-60, minimum 20, maximum 96).

Patients meeting study inclusion criteria had substantial exposure to ER/LA medications. During each patient’s qualifying period of continuous enrollment (anchored by the patient’s qualifying long-term ER/LA medication episode), study patients had a median of 1,208 days’ supply of ER/LA medications dispensed at KPW (IQR: 257-1,837, minimum 60, maximum 6,684).

Among these 3,728 patients 7% were ≤35 years of age and 26% had a diagnosis code for opioid dependence, abuse or poisoning. The distribution of study-eligible patients on these two risk factors in the study-eligible population, the sampling probabilities used to select patients for the study, and their distribution in the study sample are shown in Table 3.3.1.1.1. The sampling probabilities selected yielded a sample in which approximately half of the 2,000 patients had one or both of the risk indicators (N=996) and half did not (N=1,004).

Table 3.3.1.1.1. Distribution of Kaiser Permanente Washington (KPW) patients by opioid-related risk strata for all KPW study-eligible patients and for the 2,000 patients included in the KPW study sample according to sampling probabilities for each opioid-related risk stratum.



Opioid-related risk strata		All KPW study-eligible patients	KPW study sample (N=2,000)	
Age ≤35 years*	Diagnosis of opioid abuse, dependence, or poisoning, ever**	Number (percent)	Sampling probability	Number (percent)
Yes	Yes	108 (3%)	1	108 (6%)
Yes	No	146 (4%)	1	146 (7%)
No	Yes	869 (23%)	0.84885	742 (37%)
No	No	2605 (70%)	0.38503	1004 (50%)
Total:		3,728 (100%)	--	2,000 (100%)

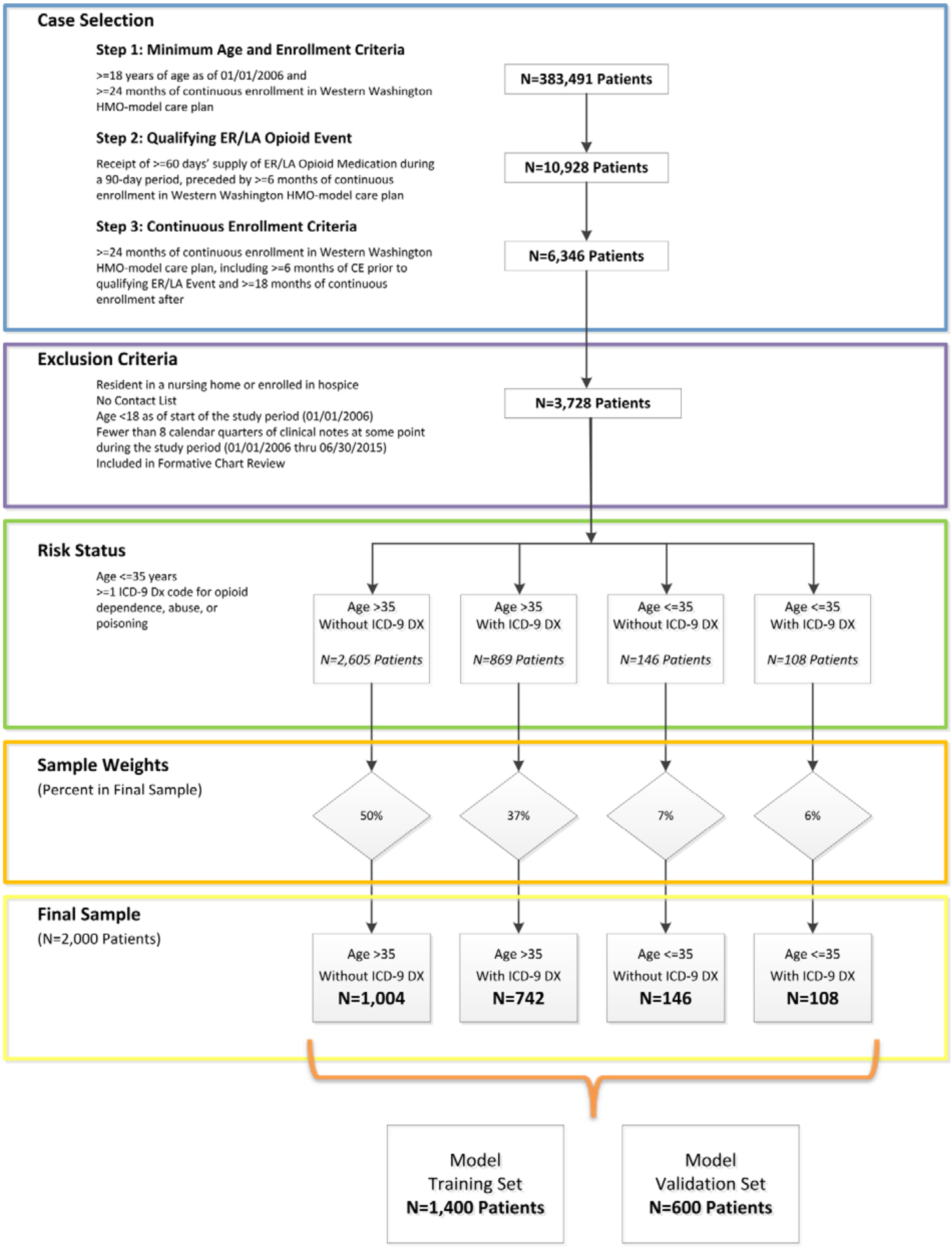
* Age as of first day of the patient's study-qualifying long-term ER/LA opioid medication episode.

** See Table 5.2 for qualifying ICD-9 diagnosis codes.

Source: Final report, pages 35-36.

A diagram summarizing the process use to select the KPW sample and the subsequent random assignment to training and validation samples is shown in Figure 3.3.1.1.2.

Figure 3.3.1.1.2. Flow diagram of Study 3B eligible population at KP Washington and process for identifying and selecting study samples.

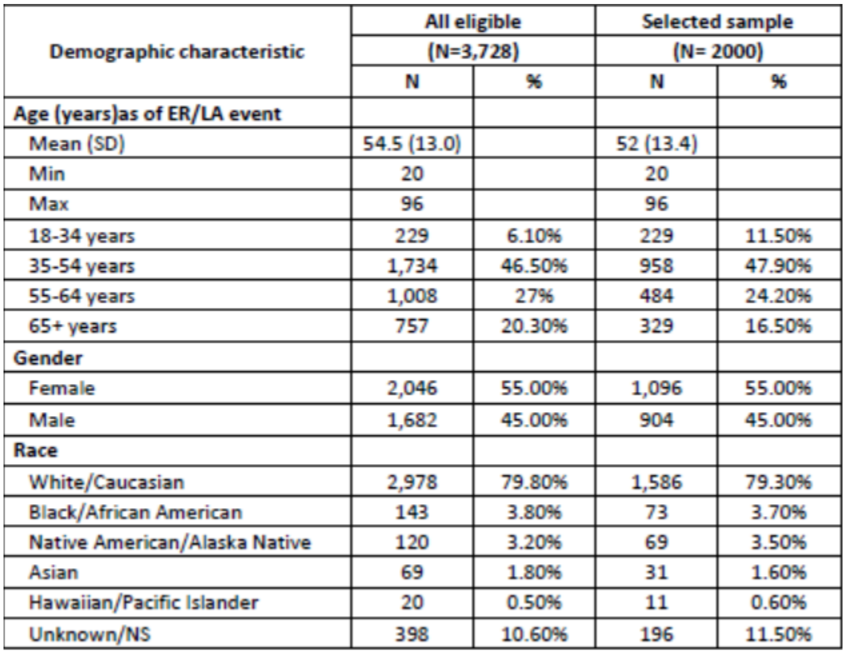

Case Selection
Step 1: Minimum Age and Enrollment Criteria
>=18 years of age as of 01/01/2006 and
>=24 months of continuous enrollment in Western Washington HMO-model care plan
N=383,491 Patients
Step 2: Qualifying ER/LA Opioid Event
Receipt of >=60 days' supply of ER/LA Opioid Medication during a 90-day period, preceded by >=6 months of continuous enrollment in Western Washington HMO-model care plan
N=10,928 Patients
Step 3: Continuous Enrollment Criteria
>=24 months of continuous enrollment in Western Washington HMO-model care plan, including >=6 months of CE prior to qualifying ER/LA Event and >=18 months of continuous enrollment after
N=6,346 Patients
Exclusion Criteria
Resident in a nursing home or enrolled in hospice
No Contact List
Age <18 as of start of the study period (01/01/2006)
Fewer than 8 calendar quarters of clinical notes at some point during the study period (01/01/2006 thru 06/30/2015)
Included in Formative Chart Review
N=3,728 Patients
Risk Status
Age <=35 years
>=1 ICD-9 Dx code for opioid dependence, abuse, or poisoning
Age >35 Without ICD-9 DX N=2,605 Patients
Age >35 With ICD-9 DX N=869 Patients
Age <=35 Without ICD-9 DX N=146 Patients
Age <=35 With ICD-9 DX N=108 Patients
Sample Weights
(Percent in Final Sample)
50%
37%
7%
6%
Final Sample
(N=2,000 Patients)
Age >35 Without ICD-9 DX N=1,004
Age >35 With ICD-9 DX N=742
Age <=35 Without ICD-9 DX N=146
Age <=35 With ICD-9 DX N=108
Model Training Set N=1,400 Patients
Model Validation Set N=600 Patients

FDA Study 3B – Group Health Research Institute

Source: Final report, page 37.

Demographic characteristics of the KPW eligible population and the 2,000 patients selected for the KPW study sample are shown in Table 3.3.1.1.3.

Table 3.3.1.1.3. Demographic characteristics of KPW patients meeting Study 3B eligibility criteria and randomly sampled for inclusion in the study cohort.

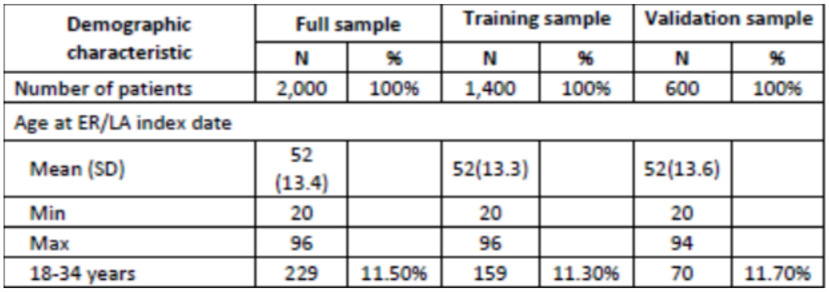


Demographic characteristic	All eligible (N=3,728)		Selected sample (N= 2000)	
	N	%	N	%
Age (years)as of ER/LA event				
Mean (SD)	54.5 (13.0)		52 (13.4)	
Min	20		20	
Max	96		96	
18-34 years	229	6.10%	229	11.50%
35-54 years	1,734	46.50%	958	47.90%
55-64 years	1,008	27%	484	24.20%
65+ years	757	20.30%	329	16.50%
Gender				
Female	2,046	55.00%	1,096	55.00%
Male	1,682	45.00%	904	45.00%
Race				
White/Caucasian	2,978	79.80%	1,586	79.30%
Black/African American	143	3.80%	73	3.70%
Native American/Alaska Native	120	3.20%	69	3.50%
Asian	69	1.80%	31	1.60%
Hawaiian/Pacific Islander	20	0.50%	11	0.60%
Unknown/NS	398	10.60%	196	11.50%

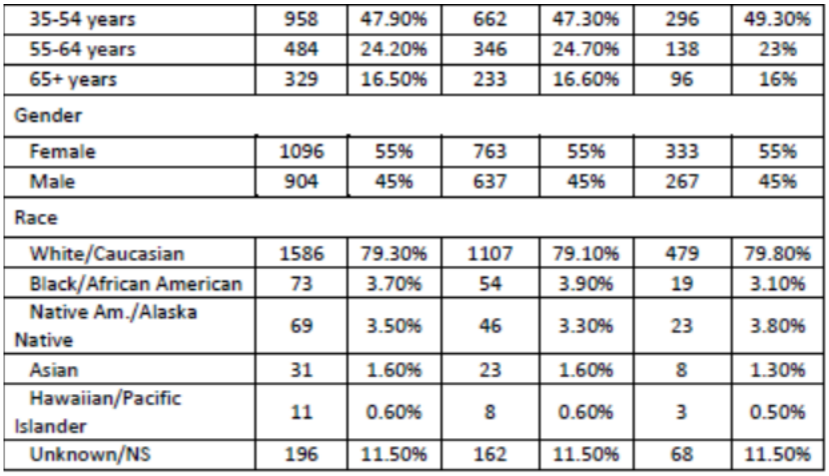
Source: Final report, page 38.

Demographic characteristics of the training and validation samples are shown in Table 3.3.1.1.4. This table shows that randomization successfully achieved the desired balance in the training and validation samples on patient demographic characteristics.

Table 3.3.1.1.4. Demographic characteristics of Kaiser Permanente Washington patients sampled for inclusion in Study 3B (N=2,000) and as randomly divided into training (N=1,400) and validation (N=600) samples.



Demographic characteristic	Full sample		Training sample		Validation sample	
	N	%	N	%	N	%
Number of patients	2,000	100%	1,400	100%	600	100%
Age at ER/LA index date						
Mean (SD)	52 (13.4)		52(13.3)		52(13.6)	
Min	20		20		20	
Max	96		96		94	
18-34 years	229	11.50%	159	11.30%	70	11.70%



35-54 years	958	47.90%	662	47.30%	296	49.30%
55-64 years	484	24.20%	346	24.70%	138	23%
65+ years	329	16.50%	233	16.60%	96	16%
Gender						
Female	1096	55%	763	55%	333	55%
Male	904	45%	637	45%	267	45%
Race						
White/Caucasian	1586	79.30%	1107	79.10%	479	79.80%
Black/African American	73	3.70%	54	3.90%	19	3.10%
Native Am./Alaska Native	69	3.50%	46	3.30%	23	3.80%
Asian	31	1.60%	23	1.60%	8	1.30%
Hawaiian/Pacific Islander	11	0.60%	8	0.60%	3	0.50%
Unknown/NS	196	11.50%	162	11.50%	68	11.50%

Source: Final report, pages 38-39.

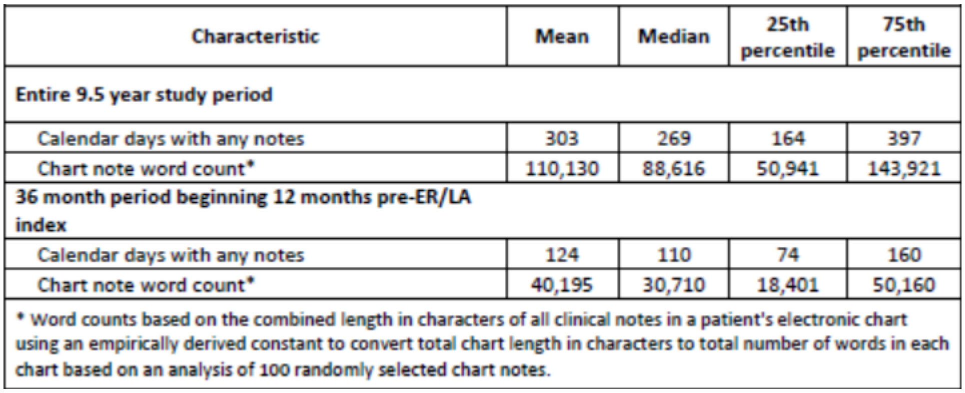
3.3.1.2 Gold Standard Chart Review Results

The electronic charts for the 2,000 patients in the KPW study cohort tended to be lengthy. As shown in Table 3.3.1.2.1, the median number of distinct calendar days with any chart notes over the 9.5-year study period was 269 (IQR: 164-397), and the median number of words per chart during the same period was >88,000 (IQR: 50,941-143,921). During the three-year period used to assemble data for the 36-month version of the AA algorithm the median number of days with chart notes was 110, and the median number of words per chart was 30,701.

The abstractor initially reviewing each chart rendered a determination regarding the presence or absence of AA with high self-rated confidence for 1,413 (71%) of the 2,000 charts. The corresponding counts for determinations rated with medium or low confidence were 548 (27%) and 39 (2%) respectively.

The 587 (29%) charts abstracted with medium or low confidence by the initial reviewer received a second, blind, independent review by one of the four abstractors not involved in its initial review, assigned to the re-review task at random. These re-reviews were independent of the initial review and independent of its results; abstractors re-reviewing chart did not know they were re-reviewing these charts. Of the 587 charts receiving additional reviews, 416 (71%) were resolved because the second reviewer’s determination regarding the patient’s AA status was concordant with that of the first reviewer. An additional 56 (10%) of the 587 charts were resolved because the second reviewer had high confidence in their determination regarding AA status (by an a priori rule, a determination made with high confidence resolved discordance between two reviewers). The remaining 115 (20%) of the 587 charts were resolved by an unblended third review by an adjudication committee which reached consensus in 80 (70%) of the 115 charts and decided by majority rule in the remaining 35 (30%) of the 115 charts. Overall, 115 (6%) of the 2,000 charts reviewed for this study were resolved by an adjudication review, and 35 charts (2% of all charts) were resolved by a non-consensus majority rule decision process.

Table 3.3.1.2.1. Characteristics of the 2,000 charts manually reviewed for PMR Observational Study 3B at KPW to assess presence/absence of prescription opioid abuse/addiction.

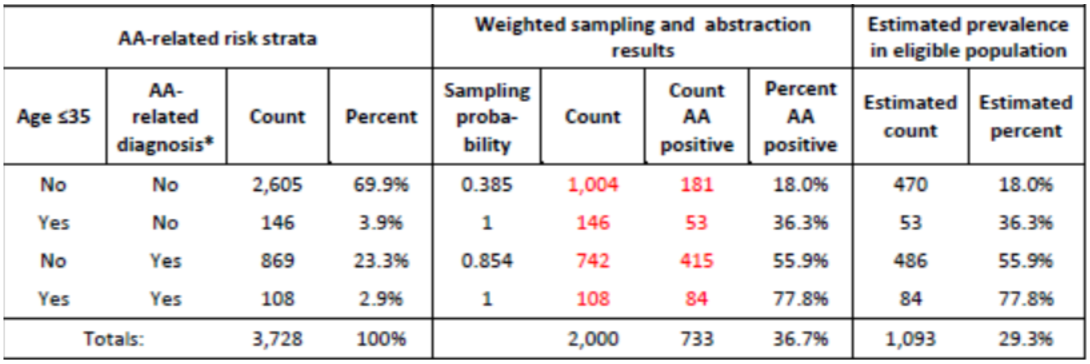


Characteristic	Mean	Median	25th percentile	75th percentile
Entire 9.5 year study period				
Calendar days with any notes	303	269	164	397
Chart note word count*	110,130	88,616	50,941	143,921
36 month period beginning 12 months pre-ER/LA index				
Calendar days with any notes	124	110	74	160
Chart note word count*	40,195	30,710	18,401	50,160

* Word counts based on the combined length in characters of all clinical notes in a patient's electronic chart using an empirically derived constant to convert total chart length in characters to total number of words in each chart based on an analysis of 100 randomly selected chart notes.

Source: Final report, page 40.

Table 3.3.1.2.2. Rates of prescription opioid abuse/addiction (AA) determined by manual chart abstraction for a weighted sample of 2,000 Kaiser Permanente Washington patients eligible for PMR Study 3B and estimated AA prevalence in the study-eligible population by risk strata.



AA-related risk strata				Weighted sampling and abstraction results				Estimated prevalence in eligible population	
Age ≤35	AA-related diagnosis*	Count	Percent	Sampling proba-bility	Count	Count AA positive	Percent AA positive	Estimated count	Estimated percent
No	No	2,605	69.9%	0.385	1,004	181	18.0%	470	18.0%
Yes	No	146	3.9%	1	146	53	36.3%	53	36.3%
No	Yes	869	23.3%	0.854	742	415	55.9%	486	55.9%
Yes	Yes	108	2.9%	1	108	84	77.8%	84	77.8%
Totals:		3,728	100%		2,000	733	36.7%	1,093	29.3%

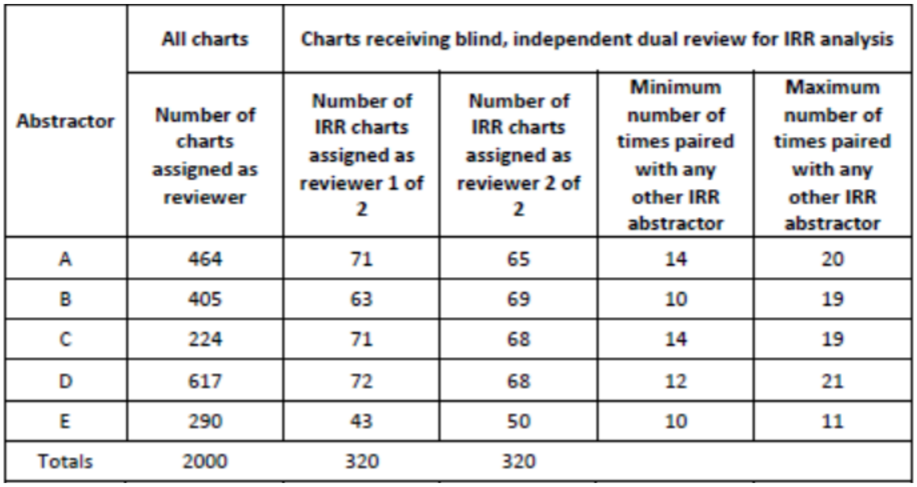
Source: Final report, page 40.

Results of the manual chart abstraction for the primary outcome of AA in the KPW study sample are shown in Table 3.3.1.2.2. A total of 733 charts were determined to be AA positive, representing 36.7% of the AA study sample, which over-sampled patients with risk factors known to be associated with AA, namely, younger age and/or ever having received an AA-related diagnosis. Table 3.3.1.2.2 also presents estimates of the prevalence of AA in the eligible KPW patient population based on a reweighting of the prevalence rates observed in the study sample. Overall, the estimated prevalence of AA among study-eligible long-term ER/LA recipients was 29.3%. The estimated prevalence was 18.0% among patients who did not qualify for either AA-related risk stratum and was 77.8% among patients who qualified for both AA-risk related strata. Estimated AA prevalence for patients 35 years of age or younger was 36.3%, and 55.9% among patients ever receiving an AA-related diagnosis.

Inter-rater reliability review results

Before any of the 2,000 study charts were reviewed investigators selected a random subset of 320 charts to receive blind, independent, dual review for purposes of assessing IRR. Two chart abstractors were randomly assigned to review each IRR chart. As shown in Table 3.3.1.2.3, abstractor assignment assured that each abstractor was paired at least 10 times with each of the other four abstractors, and that each abstractor's most frequently occurring pairing with another abstractor occurred no more than twice as frequently as their least frequently occurring pairing with another abstractor.

Table 3.3.1.2.3. Summary of all charts assigned and charts assigned for blind, independent, dual review for purposes of assessing inter-rater reliability (IRR) by abstractor.



Abstractor	All charts	Charts receiving blind, independent dual review for IRR analysis			
	Number of charts assigned as reviewer	Number of IRR charts assigned as reviewer 1 of 2	Number of IRR charts assigned as reviewer 2 of 2	Minimum number of times paired with any other IRR abstractor	Maximum number of times paired with any other IRR abstractor
A	464	71	65	14	20
B	405	63	69	10	19
C	224	71	68	14	19
D	617	72	68	12	21
E	290	43	50	10	11
Totals	2000	320	320		

Consistent with chart abstraction rules defined in advance in the chart abstraction protocol (available as a separate document), the 1,413 charts reviewed with high confidence by their initial reviewers received no further review. Investigators implemented this rule to improve efficiency and used the IRR chart reviews to assess whether re-reviewing high-confidence charts would have resulted in non-trivial differences with respect to the primary study outcome—AA presence or absence.

Of the 320 charts assigned to the IRR review set 235 (73%) were among the 1,413 (71%) charts receiving a high-confidence review by the initial reviewer. Based on results from these paired, blind, independent reviews regarding AA status (AA positive or AA negative) the Cohen's kappa coefficient was 0.83. There are no widely accepted rules for what magnitude of kappa corresponds to adequate agreement. Nevertheless, Landis and Koch characterized values 0.21–0.40 as fair, 0.41–0.60 as moderate, 0.61–0.80 as substantial, and 0.81–1 as almost perfect agreement. Fleiss characterized kappa statistics >0.75 as excellent.

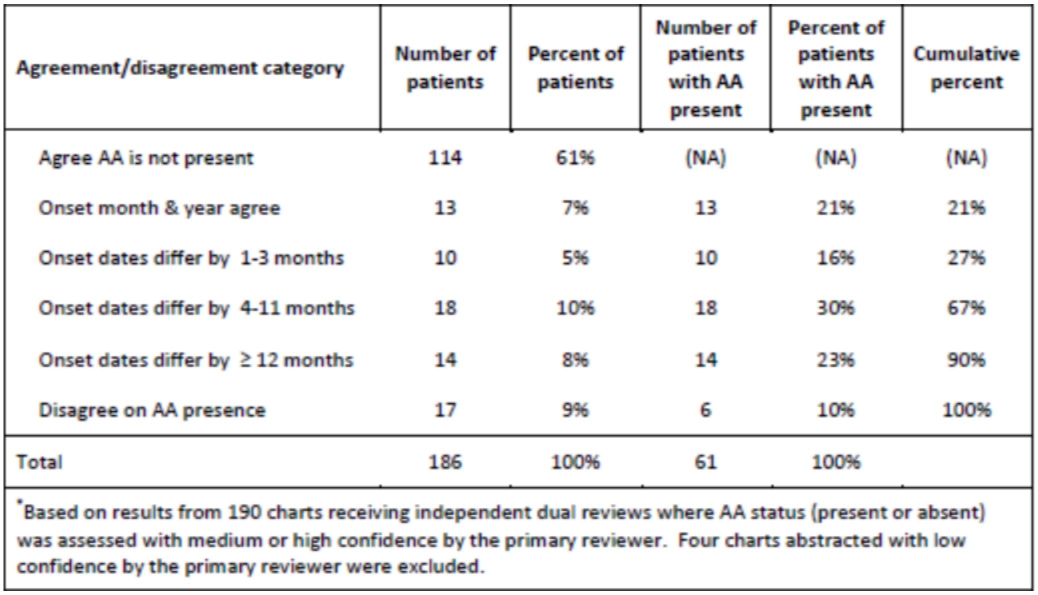
Reviewer Comment:

This approach to chart adjudication appears reasonable and is designed to minimize the potential for bias associated with an outlier reviewer's perspective. The kappa statistic of >0.8 indicates a high level of inter-rater agreement among abstractors.

Inter-rater agreement regarding determination of AA onset

As shown in Table 3.3.1.2.4, among charts selected for IRR review, chart abstractors agreed regarding the month and year of AA onset for 21% of charts and agreed within a three-month time window for an additional 16% of charts. Thirty percent of chart were assigned onset dates that differed by 4-11 months, and an additional 23% were discordant by 12 or more months.

Table 3.3.1.2.4. Inter-rater agreement regarding onset date of prescription opioid abuse/addiction (AA) among abstractors independently reviewing the same patient chart.*



Agreement/disagreement category	Number of patients	Percent of patients	Number of patients with AA present	Percent of patients with AA present	Cumulative percent
Agree AA is not present	114	61%	(NA)	(NA)	(NA)
Onset month & year agree	13	7%	13	21%	21%
Onset dates differ by 1-3 months	10	5%	10	16%	27%
Onset dates differ by 4-11 months	18	10%	18	30%	67%
Onset dates differ by ≥ 12 months	14	8%	14	23%	90%
Disagree on AA presence	17	9%	6	10%	100%
Total	186	100%	61	100%	

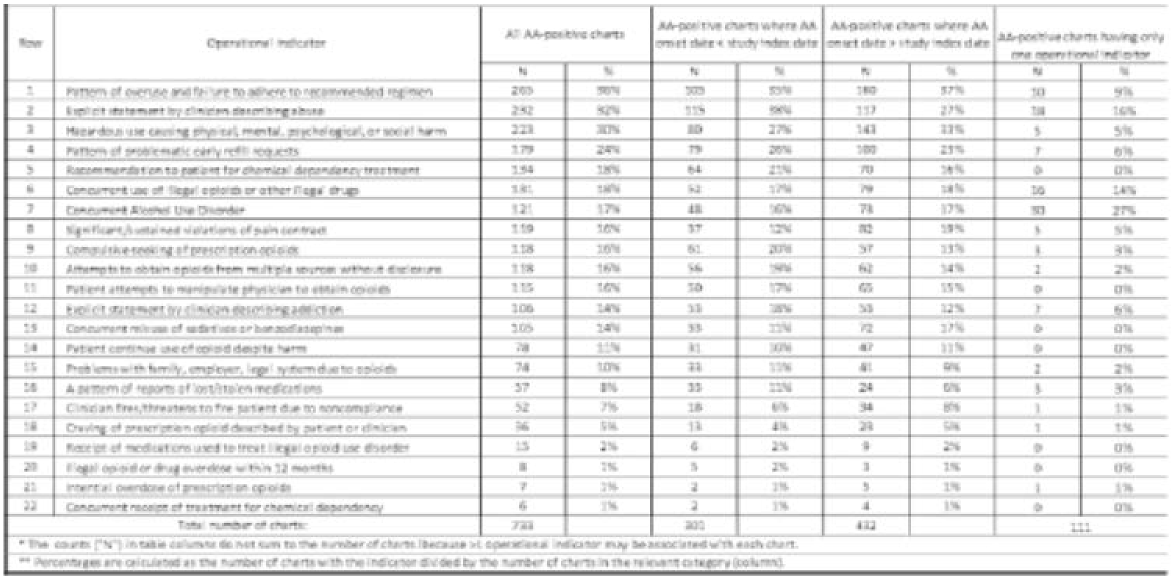
*Based on results from 190 charts receiving independent dual reviews where AA status (present or absent) was assessed with medium or high confidence by the primary reviewer. Four charts abstracted with low confidence by the primary reviewer were excluded.

Operational indicators of AA

Operational indicators of AA documented in the chart and recorded by the initial reviewer of each chart are summarized in Table 3.3.1.2.5 (see page 44 of final report, column "All"). As shown, the most frequently recorded indicator was "patterns of overuse and failure to adhere to recommended regimen," appearing in 36% of charts. Two other indicators were also recorded for ≥30% of charts: "Explicit statement by clinician describing abuse" and "Hazardous use causing physical, mental, psychological, or social harm." The number of operational indicators identified by manual chart abstraction for each of the 733 AA-positive charts ranged from 1 to 14. Twenty percent of AA-positive chart reviews identified one operational indicator, 22% identified two, 20% identified 3, 18% identified 4, and 20% identified 5-14 operational indicators. These categories were mutually exclusive. It should be noted that, as directed by the chart abstraction protocol, chart abstractors were to record at least one operational indicator for each chart in which they determined evidence of AA was present. Abstractors were not instructed to identify and record all operational indicators recorded in AA positive charts. Therefore, care must be taken in interpreting the data in Table 3.3.1.2.5.

Table 3.3.1.2.5. Number* and percent of study 3B charts where prescription opioid abuse/addiction (AA) was identified by manual abstraction, for all AA-positive charts, AA-positive where the AA onset date preceded the study index date,**

AA-positive charts where the AA onset date followed the study index date, and for AA-positive charts where only one operational indicator was identified by the chart abstractor.



Row	Operational Indicator	All AA-positive charts		AA-positive charts where AA onset date < study index date		AA-positive charts where AA onset date > study index date		AA-positive charts having only one operational indicator	
		N	%	N	%	N	%	N	%
1	Pattern of overuse and failure to adhere to recommended regimen	265	36%	105	35%	160	37%	10	9%
2	Explicit statement by clinician describing abuse	232	32%	115	38%	117	27%	18	16%
3	Hazardous use causing physical, mental, psychological, or social harm	223	30%	80	27%	143	33%	5	5%
4	Pattern of problematic early refill requests	179	24%	79	26%	100	23%	7	6%
5	Recommendation to patient for chemical dependency treatment	134	18%	64	21%	70	16%	0	0%
6	Concurrent use of illegal opioids or other illegal drugs	141	19%	62	21%	79	18%	16	14%
7	Concurrent Alcohol Use Disorder	121	17%	48	16%	73	17%	30	27%
8	Significant/sustained violations of pain contract	119	16%	37	12%	82	19%	5	5%
9	Compulsive seeking of prescription opioids	118	16%	61	20%	57	13%	3	3%
10	Attempts to obtain opioids from multiple sources without disclosure	118	16%	56	19%	62	14%	2	2%
11	Patient attempts to manipulate physician to obtain opioids	115	16%	50	17%	65	15%	0	0%
12	Explicit statement by clinician describing addiction	106	14%	53	18%	53	12%	7	6%
13	Concurrent misuse of sedatives or benzodiazepines	105	14%	33	11%	72	17%	0	0%
14	Patient continue use of opioid despite harm	78	11%	31	10%	47	11%	0	0%
15	Problems with family, employer, legal system due to opioids	74	10%	33	11%	41	9%	2	2%
16	A pattern of reports of lost/stolen medications	57	8%	33	11%	24	6%	3	3%
17	Clinician fires/threatens to fire patient due to noncompliance	52	7%	18	6%	34	8%	1	1%
18	Craving of prescription opioid described by patient or clinician	36	5%	13	4%	23	5%	1	1%
19	Receipt of medications used to treat illegal opioid use disorder	15	2%	6	2%	9	2%	0	0%
20	Illegal opioid or drug overdose within 12 months	8	1%	5	2%	3	1%	0	0%
21	Intentional overdose of prescription opioids	7	1%	2	1%	5	1%	1	1%
22	Concurrent receipt of treatment for chemical dependency	6	1%	2	1%	4	1%	0	0%
	Total number of charts:	733		301		432		111	

* The counts ("N") in table columns do not sum to the number of charts because >1 operational indicator may be associated with each chart.

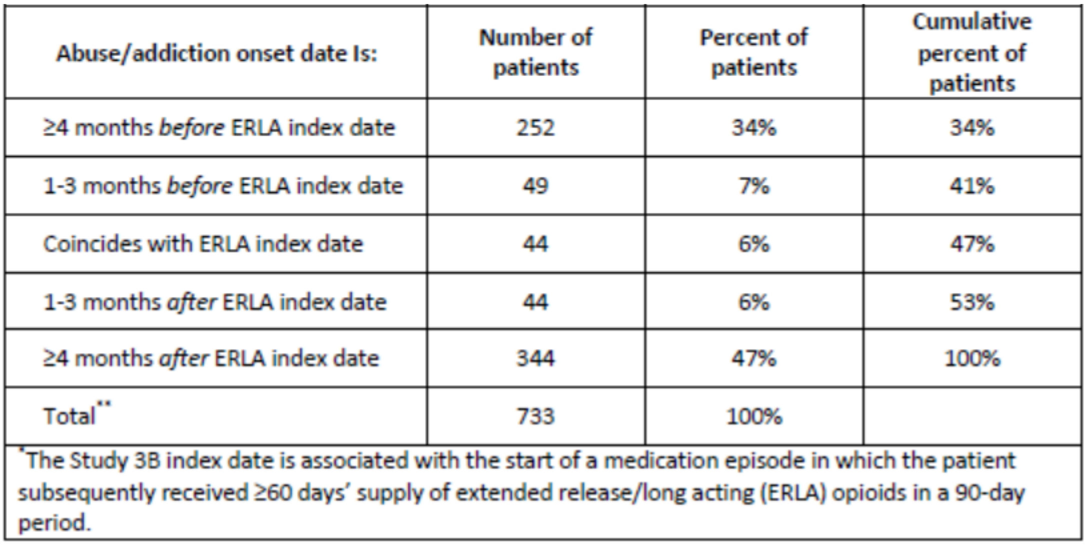
** Percentages are calculated as the number of charts with the indicator divided by the number of charts in the relevant category (column).

Source: Final report, page 44.

Temporal ordering of study index date and AA onset

As shown in Table 3.3.1.2.6, the onset date in a chart containing evidence of AA often preceded the Study 3B ER/LA index date. For 41% of AA positive charts the onset date was before the index date, and in an additional 6% of charts the onset date and index date coincided.

Table 3.3.1.2.6. Distribution of patients with gold standard determinations of prescription opioid abuse/addiction (AA) by temporal proximity of the abstractor-determined AA onset date relative to each patient's long-term ERLA opioid analgesic index date.*



Abuse/addiction onset date Is:	Number of patients	Percent of patients	Cumulative percent of patients
≥4 months *before* ERLA index date	252	34%	34%
1-3 months *before* ERLA index date	49	7%	41%
Coincides with ERLA index date	44	6%	47%
1-3 months *after* ERLA index date	44	6%	53%
≥4 months *after* ERLA index date	344	47%	100%
Total**	733	100%	

*The Study 3B index date is associated with the start of a medication episode in which the patient subsequently received ≥60 days' supply of extended release/long acting (ERLA) opioids in a 90-day period.

Source: Final report, page 45.

Reviewer Comment:

It is notable that a substantial proportion of patients demonstrated evidence of AA before receiving an ER/LA opioid analgesic. One possibility is that clinicians may be "channeling" patients exhibiting AA behaviors from immediate-release opioid analgesics to ER/LA opioid analgesics. Some prescribers may view ER/LA opioid analgesics' steady release of active pharmaceutical ingredient (API) as beneficial to avoid the reinforcing peak-and-trough psychotropic effects observed with immediate-release opioid analgesics. However, this presumed pharmacokinetic benefit is undercut by the fact that ER/LA opioid analgesics typically contain more API than immediate-release opioid analgesics: patients demonstrating AA behaviors who are channeled to ER/LA opioid analgesics on this basis therefore may experience a reinforcement of the AA behavior based on the increased total API exposure. These results indicate that there may be value in performing additional research on the complex temporal relationship between ER/LA opioid analgesic exposure, AA behavior, and possible prescriber channeling.

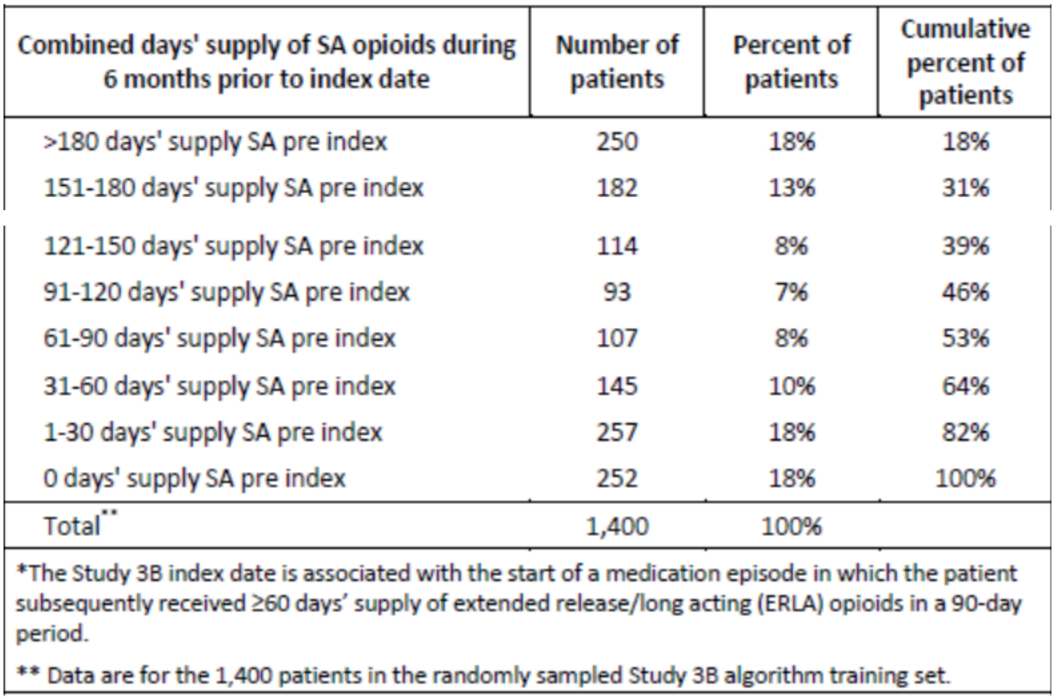
It is also notable that the age distribution in this sample is skewed towards the older adult population. KP has stringent opioid dispensing populations, and ER/LA opioid analgesics represent high-risk products because they contain high doses of active pharmaceutical ingredient. One possibility is that individuals in the KP system receiving ER/LA opioid analgesics may have been older and sicker than the investigators anticipated, as age and medical complexity often correlate.

Exposure to opioids prior to study index date

Patients in the study cohort were highly likely to be exposed to substantial amounts of short- acting (SA) prescription opioids prior to their Study 3B index date. As shown in Table 3.3.1.2.7, over half (53%) of patients in the 1,400 patient training had been exposed to at least 60 days' supply of SA opioids in the six month period preceding the Study 3B index date, and almost one-third (31%) had received more than 150 days' supply.

Eighteen percent of patients did not receive any SA opioids in the six months prior to their index date.

Table 3.3.1.2.7. Distribution of Study 3B training sample patients by combined days' supply of short acting (SA) prescription opioids dispensed in the 6 months prior to each patient's study index date*



Combined days' supply of SA opioids during 6 months prior to index date	Number of patients	Percent of patients	Cumulative percent of patients
>180 days' supply SA pre index	250	18%	18%
151-180 days' supply SA pre index	182	13%	31%
121-150 days' supply SA pre index	114	8%	39%
91-120 days' supply SA pre index	93	7%	46%
61-90 days' supply SA pre index	107	8%	53%
31-60 days' supply SA pre index	145	10%	64%
1-30 days' supply SA pre index	257	18%	82%
0 days' supply SA pre index	252	18%	100%
Total**	1,400	100%	

*The Study 3B index date is associated with the start of a medication episode in which the patient subsequently received ≥60 days' supply of extended release/long acting (ERLA) opioids in a 90-day period.

** Data are for the 1,400 patients in the randomly sampled Study 3B algorithm training set.

Source: Final report, page 45.

Claims-based AA classification algorithm

These are the results of the five AA algorithms developed and evaluated using data from the primary study site. There were four algorithms based entirely on claims (Sentinel) data and one algorithm that incorporated additional data extracted from the EHR. The five AA classification algorithms are:

1) Full-period claims-based algorithm
2) 36-month claims-based algorithm
3) Full-period simple ICD-9 code algorithm
4) 36-month simple ICD-9 code algorithm
5) 36-month EHR-enhanced algorithm

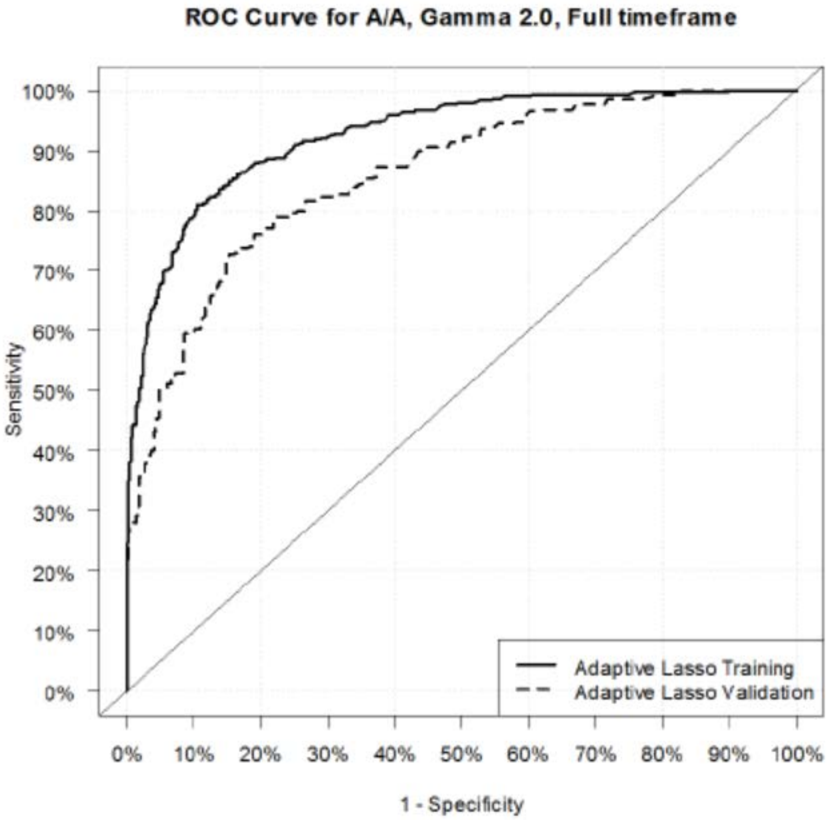
Investigators operationalized potential predictors of AA based on information available from medical claims data, including diagnoses, medications, encounters, and procedures. A high-level summary of the categories for which potential predictors were operationalized are included as an appendix in the final report. Based on the results of examining distributions of a large number of potential predictor variables by patients' AA status in training data (AA positive versus AA negative), investigators selected potential predictors with high face validity or strong empirical associations with AA for further development. Investigators then created several versions of a various potential predictor variables. This resulted in a large number of operationalized potential predictors for both the full-term model and the 36-month model. For the 36-month model, investigators operationalized a total of 1,122 potential predictors, all of which were included in the adaptive LASSO variable selection process to identify the subset of predictors which, as a group, classified patients with respect to AA status with the best performance characteristics. These 1,122 potential predictors are described in Appendix 3 of the final report.

Full-period claims-based algorithm

Results for the full-term AA classification algorithm, based on up to 9.5 years of data for patients from KPW data are shown in Figure 3.3.1.2.8 and Table 3.3.1.2.8. The investigators assessed how the algorithm performed at different risk score cut-points. As the figure shows, the algorithm achieved fairly good performance in the training data, but performance degraded substantially when applied to previously unseen validation data, suggesting overfitting issues. The magnitude of overfitting is revealed in row 10 of Table 3.3.1.2.8, where the balancing point[2] for sensitivity and PPV in the training data reached sensitivity of 0.781 and PPV of 0.781, but dropped substantially in the validation data, where they reached 0.638 and 0.692, respectively. This level of performance is below the threshold of minimally acceptable performance, which investigators defined as sensitivity ≥0.750 and PPV ≥0.750. Investigators also note that the full-term model is only relevant in settings such as KPW where average patient follow-up is much longer than it is in typical commercial insurance settings.

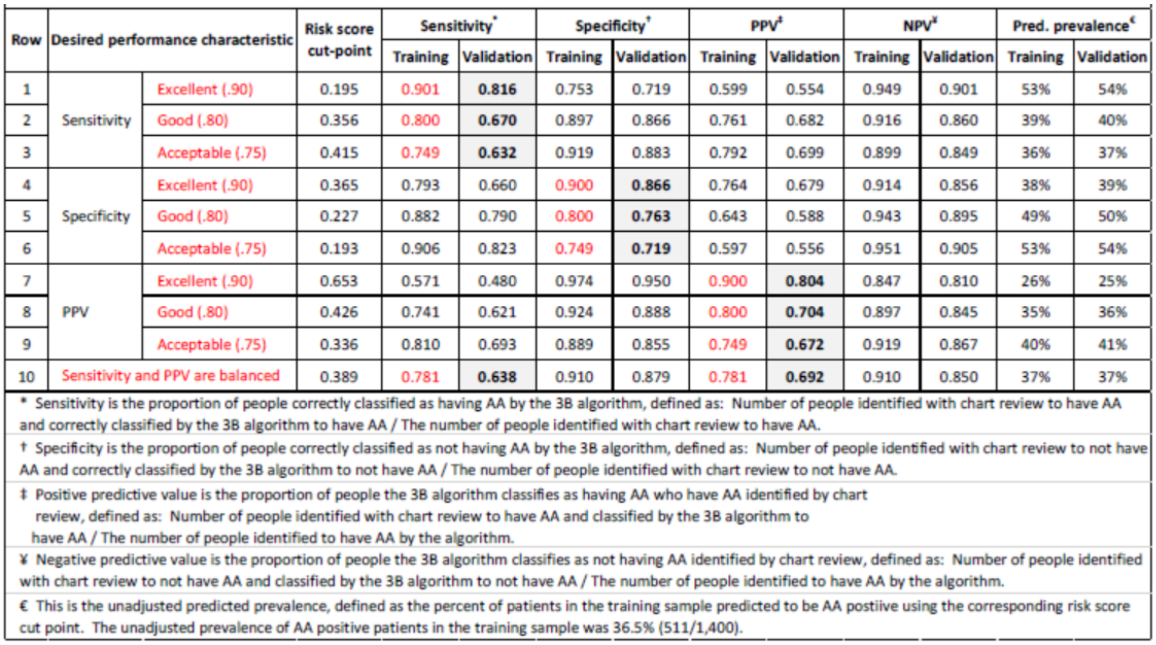
Figure 3.3.1.2.8. ROC curve for the full-period claims-based algorithm.

[2] For a given population’s ROC curve, sliding the risk score cut-off point along the curve changes test performance metrics: sensitivity, specificity, PPV, and NPV. The investigators created an analytic model that slides the risk score cut-off point along the ROC, and calculates the cut-off point that simultaneously maximizes/optimizes PPV and sensitivity for that population. This approach adds value to the algorithm by not simply identifying the most obvious cases of AA (high PPV) at the expense of missing cases that are less obvious (low sensitivity).



Source: Final report, page 50.

Table 3.3.1.2.8. Full-term AA classification algorithm performance characteristics (sensitivity, specificity, PPV, NPV, percent predicted AA positive) in KPW training and validation samples for cut points of the AA risk score with desired performance characteristics selected based on training data.



Row	Desired performance characteristic		Risk score cut-point	Sensitivity*		Specificity†		PPV‡		NPV¥		Pred. prevalence€	
				Training	Validation	Training	Validation	Training	Validation	Training	Validation	Training	Validation
1	Sensitivity	Excellent (.90)	0.195	0.901	**0.816**	0.753	0.719	0.599	0.554	0.949	0.901	53%	54%
2		Good (.80)	0.356	0.800	**0.670**	0.897	0.866	0.761	0.682	0.916	0.860	39%	40%
3		Acceptable (.75)	0.415	0.749	**0.632**	0.919	0.883	0.792	0.699	0.899	0.849	36%	37%
4	Specificity	Excellent (.90)	0.365	0.793	0.660	0.900	**0.866**	0.764	0.679	0.914	0.856	38%	39%
5		Good (.80)	0.227	0.882	0.790	0.800	**0.763**	0.643	0.588	0.943	0.895	49%	50%
6		Acceptable (.75)	0.193	0.906	0.823	0.749	**0.719**	0.597	0.556	0.951	0.905	53%	54%
7	PPV	Excellent (.90)	0.653	0.571	0.480	0.974	0.950	0.900	**0.804**	0.847	0.810	26%	25%
8		Good (.80)	0.426	0.741	0.621	0.924	0.888	0.800	**0.704**	0.897	0.845	35%	36%
9		Acceptable (.75)	0.336	0.810	0.693	0.889	0.855	0.749	**0.672**	0.919	0.867	40%	41%
10	Sensitivity and PPV are balanced		0.389	0.781	**0.638**	0.910	0.879	0.781	**0.692**	0.910	0.850	37%	37%

* Sensitivity is the proportion of people correctly classified as having AA by the 3B algorithm, defined as: Number of people identified with chart review to have AA and correctly classified by the 3B algorithm to have AA / The number of people identified with chart review to have AA.

† Specificity is the proportion of people correctly classified as not having AA by the 3B algorithm, defined as: Number of people identified with chart review to not have AA and correctly classified by the 3B algorithm to not have AA / The number of people identified with chart review to not have AA.

‡ Positive predictive value is the proportion of people the 3B algorithm classifies as having AA who have AA identified by chart review, defined as: Number of people identified with chart review to have AA and classified by the 3B algorithm to have AA / The number of people identified to have AA by the algorithm.

¥ Negative predictive value is the proportion of people the 3B algorithm classifies as not having AA identified by chart review, defined as: Number of people identified with chart review to not have AA and classified by the 3B algorithm to not have AA / The number of people identified to have AA by the algorithm.

€ This is the unadjusted predicted prevalence, defined as the percent of patients in the training sample predicted to be AA postiive using the corresponding risk score cut point. The unadjusted prevalence of AA positive patients in the training sample was 36.5% (511/1,400).

Source: Final report, page 51.

Reviewer Comment:
The full term claims algorithm demonstrated a sensitivity of 0.64 and a PPV of 0.69 in the KPW validation data at the balancing point, both of which fall below the acceptable levels of performance established a priori.

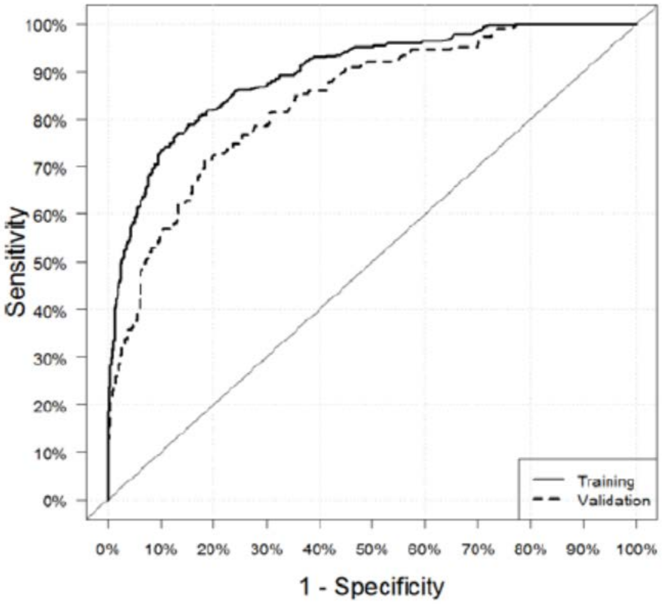
It is notable that the algorithm's specificity is relatively robust at the balancing point in both the training and validation data sets.

36-month claims-based algorithm

The adaptive LASSO variable selection processes yielded a set of 53 predictors in the 36-month claims-based algorithm. Predictors included age, gender, diagnosis of opioid dependence; diagnoses of comorbidities such as mental health disorders, alcohol use disorder, non-opioid drug dependence, tobacco use disorder and anxiety disorder; various measures of opioid dispensings based on days' supply and morphine equivalent dose; dispensing of opioids concomitantly with other medications such as benzodiazepines; various measures of early refills of opioid medications; opioid dispensings in emergency room and/or urgent care settings, history of receiving medications used to treat drug dependence, coincidence of urine drug screening procedures and dispensings of opioid medications; measures based on pain diagnoses, and interaction terms based on patient age. The complete list of these 53 predictors, along with their coefficients and the formulation used to calculate the AA risk score, is presented in Appendix 4 of the final report.

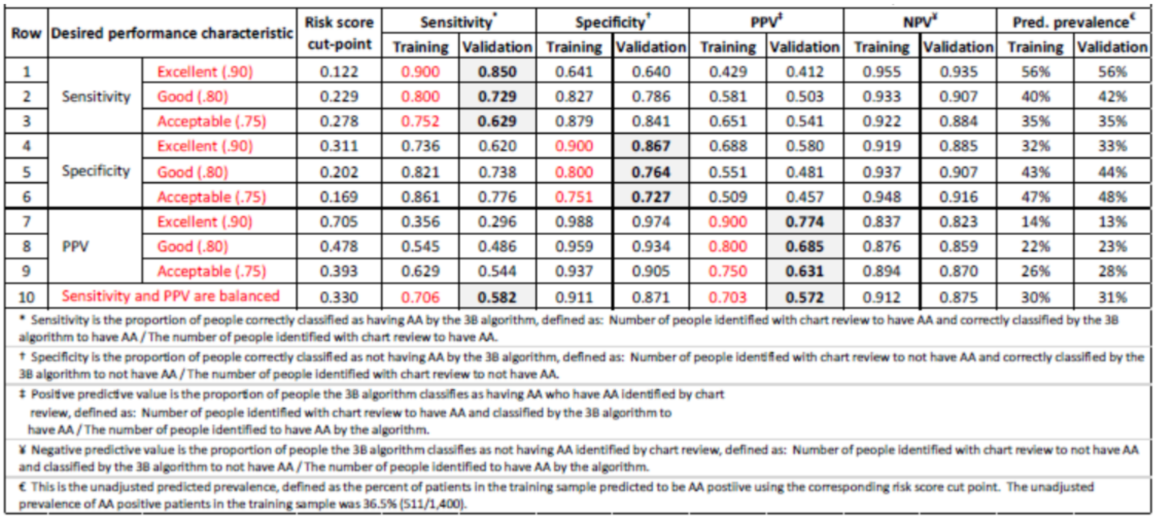
Results for the 36-month version of the AA classification algorithm, based on up to 36 months of data for each KPW patient, are shown in Figure 3.3.1.2.9 and Table 3.3.1.2.9. The ROC curves and the performance metrics summarized in the table show that the 36-month model did not perform quite as well as the full-term model, though the results are not drastically different. Comparing the algorithm's performance where sensitivity and PPV are approximately balanced (row 10 in Tables 3.3.1.2.8 and 3.3.1.2.9) the performance in validation data for sensitivity and PPV were 0.638 and 0.692 respectively in the full-term version of the model, compared to 0.582 and 0.572, respectively, in the 36-month version of the model.

Figure 3.3.1.2.9. ROC curve for the 36-month claims-based algorithm.



Source: Final report, page 52.

Table 3.3.1.2.9. 36-month AA classification algorithm performance characteristics (sensitivity, specificity, PPV, NPV, percent predicted AA positive) in KPW training and validation samples for cut points of the AA risk score with desired performance characteristics selected based on training data.



Row	Desired performance characteristic		Risk score cut-point	Sensitivity[*]		Specificity[†]		PPV[‡]		NPV[¥]		Pred. prevalence[€]	
				Training	Validation	Training	Validation	Training	Validation	Training	Validation	Training	Validation
1	Sensitivity	Excellent (.90)	0.122	0.900	**0.850**	0.641	0.640	0.429	0.412	0.955	0.935	56%	56%
2		Good (.80)	0.229	0.800	**0.729**	0.827	0.786	0.581	0.503	0.933	0.907	40%	42%
3		Acceptable (.75)	0.278	0.752	**0.629**	0.879	0.841	0.651	0.541	0.922	0.884	35%	35%
4	Specificity	Excellent (.90)	0.311	0.736	0.620	0.900	**0.867**	0.688	0.580	0.919	0.885	32%	33%
5		Good (.80)	0.202	0.821	0.738	0.800	**0.764**	0.551	0.481	0.937	0.907	43%	44%
6		Acceptable (.75)	0.169	0.861	0.776	0.751	**0.727**	0.509	0.457	0.948	0.916	47%	48%
7	PPV	Excellent (.90)	0.705	0.356	0.296	0.988	0.974	0.900	**0.774**	0.837	0.823	14%	13%
8		Good (.80)	0.478	0.545	0.486	0.959	0.934	0.800	**0.685**	0.876	0.859	22%	23%
9		Acceptable (.75)	0.393	0.629	0.544	0.937	0.905	0.750	**0.631**	0.894	0.870	26%	28%
10	Sensitivity and PPV are balanced		0.330	0.706	**0.582**	0.911	0.871	0.703	**0.572**	0.912	0.875	30%	31%

[*] Sensitivity is the proportion of people correctly classified as having AA by the 3B algorithm, defined as: Number of people identified with chart review to have AA and correctly classified by the 3B algorithm to have AA / The number of people identified with chart review to have AA.

[†] Specificity is the proportion of people correctly classified as not having AA by the 3B algorithm, defined as: Number of people identified with chart review to not have AA and correctly classified by the 3B algorithm to not have AA / The number of people identified with chart review to not have AA.

[‡] Positive predictive value is the proportion of people the 3B algorithm classifies as having AA who have AA identified by chart review, defined as: Number of people identified with chart review to have AA and classified by the 3B algorithm to have AA / The number of people identified to have AA by the algorithm.

[¥] Negative predictive value is the proportion of people the 3B algorithm classifies as not having AA identified by chart review, defined as: Number of people identified with chart review to not have AA and classified by the 3B algorithm to not have AA / The number of people identified to have AA by the algorithm.

[€] This is the unadjusted predicted prevalence, defined as the percent of patients in the training sample predicted to be AA postiive using the corresponding risk score cut point. The unadjusted prevalence of AA positive patients in the training sample was 36.5% (511/1,400).

Source: Final report, page 53.

Reviewer Comment:
The 36-month claims algorithm demonstrated a sensitivity of 0.58 and a PPV of 0.57 in the KPW validation data at the balancing point, both of which fall below the acceptable levels of performance established a priori.

It is notable that the algorithm's specificity is relatively robust at the balancing point in both the training and validation data sets.

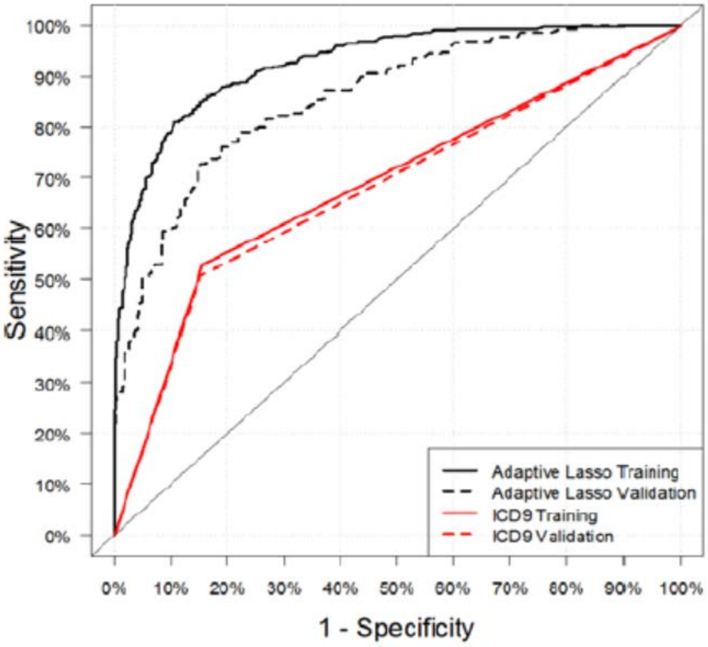
Simple ICD-9 code algorithms: Full-period and 36-month versions

The simple ICD-9 code algorithms performed markedly less well in either its full-term or 36-month versions. When interpreting the ROC curves for these simple algorithms it should be noted that the algorithm yields only a single point in the two-dimensional sensitivity/specificity plane. This is because, as described above, the simple model yields a binary prediction for each patient—not an interval level risk score that can be assessed at varying cut points. Though the ROC curves present the results with lines, only the inflection point of the line is meaningful.

Full-period simple ICD-9 code algorithm

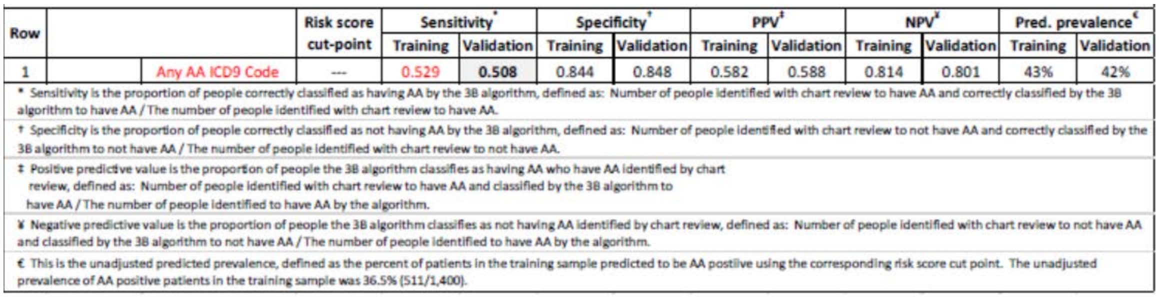
The full-period simple ICD-9 code algorithm achieved sensitivity of 0.508 and PPV of 0.588 in validation data (Table 3.3.1.2.10). As illustrated in the ROC curve (Figure 3.3.1.2.10) this is substantially below the performance of the full-period claims-based algorithm.

Figure 3.3.1.2.10. ROC curve for the full-period simple ICD-9 code algorithm.



Source: Final report, page 54.

Table 3.3.1.2.10. Full-term simple ICD-9 AA classification algorithm performance characteristics (sensitivity, specificity, PPV, NPV, percent predicted AA positive) in KPW training and validation samples for cut points of the AA risk score with desired performance characteristics selected based on training data.



Row			Risk score cut-point	Sensitivity*		Specificity†		PPV‡		NPV¥		Pred. prevalence€	
				Training	Validation	Training	Validation	Training	Validation	Training	Validation	Training	Validation
1		Any AA ICD9 Code	---	0.529	**0.508**	0.844	0.848	0.582	0.588	0.814	0.801	43%	42%

* Sensitivity is the proportion of people correctly classified as having AA by the 3B algorithm, defined as: Number of people identified with chart review to have AA and correctly classified by the 3B algorithm to have AA / The number of people identified with chart review to have AA.

† Specificity is the proportion of people correctly classified as not having AA by the 3B algorithm, defined as: Number of people identified with chart review to not have AA and correctly classified by the 3B algorithm to not have AA / The number of people identified with chart review to not have AA.

‡ Positive predictive value is the proportion of people the 3B algorithm classifies as having AA who have AA identified by chart review, defined as: Number of people identified with chart review to have AA and classified by the 3B algorithm to have AA / The number of people identified to have AA by the algorithm.

¥ Negative predictive value is the proportion of people the 3B algorithm classifies as not having AA identified by chart review, defined as: Number of people identified with chart review to not have AA and classified by the 3B algorithm to not have AA / The number of people identified to have AA by the algorithm.

€ This is the unadjusted predicted prevalence, defined as the percent of patients in the training sample predicted to be AA postive using the corresponding risk score cut point. The unadjusted prevalence of AA positive patients in the training sample was 36.5% (511/1,400).

Source: Final report, page 55.

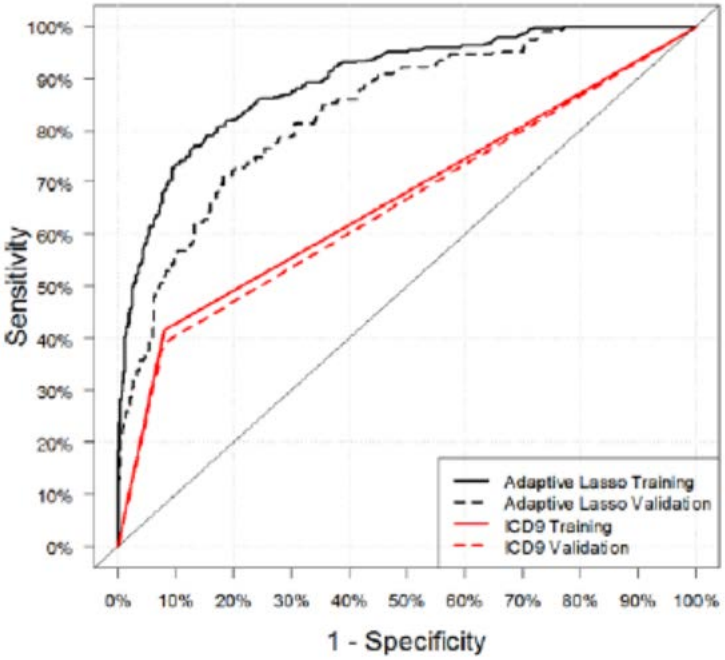
Reviewer Comment:
The full period simple ICD-9 code algorithm demonstrated a sensitivity of 0.51 and a PPV of 0.59 in the KPW validation data, both of which fall below the acceptable levels of performance established a priori.

It is notable that the algorithm’s specificity is relatively robust at the balancing point in both the training and validation data sets.

36-month ICD-9 code algorithm

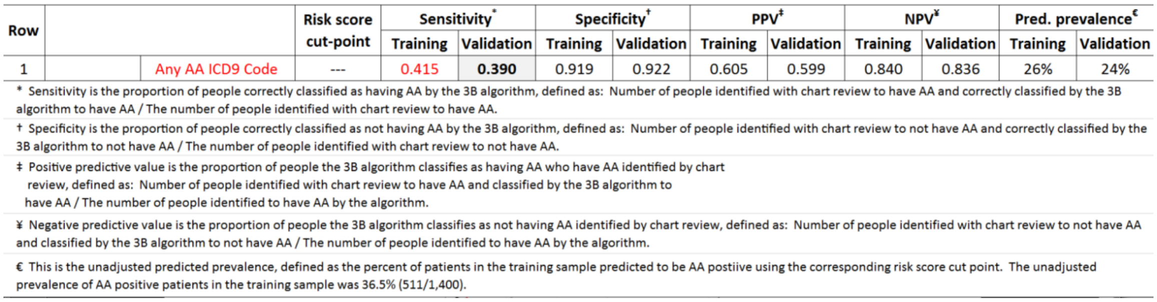
The 36-month simple ICD-9 code algorithm achieved sensitivity of 0.390 and PPV of 0.599 in validation data (Table 3.3.1.2.11). As illustrated in the ROC curve (Figure 3.3.1.2.11) this is below the performance of the 36-month claims based algorithm. However, this performance approaches that of the claims-based algorithm at that particular level of sensitivity and specificity, as indicated in the ROC curve. These results suggest the simple algorithm has much better specificity (0.922) than sensitivity (0.390) in validation data.

Figure 3.3.1.2.11. ROC curve for the 36-month simple ICD-9 code algorithm.



Source: Final report, page 56.

Table 3.3.1.2.11. 36-month simple ICD-9 AA classification algorithm performance characteristics (sensitivity, specificity, PPV, NPV, percent predicted AA positive) in KPW training and validation samples for cut points of the AA risk score with desired performance characteristics selected based on training data.



Row			Risk score	Sensitivity*		Specificity†		PPV‡		NPV¥		Pred. prevalence€	
			cut-point	Training	Validation	Training	Validation	Training	Validation	Training	Validation	Training	Validation
1		Any AA ICD9 Code	---	0.415	**0.390**	0.919	0.922	0.605	0.599	0.840	0.836	26%	24%

* Sensitivity is the proportion of people correctly classified as having AA by the 3B algorithm, defined as: Number of people identified with chart review to have AA and correctly classified by the 3B algorithm to have AA / The number of people identified with chart review to have AA.

† Specificity is the proportion of people correctly classified as not having AA by the 3B algorithm, defined as: Number of people identified with chart review to not have AA and correctly classified by the 3B algorithm to not have AA / The number of people identified with chart review to not have AA.

‡ Positive predictive value is the proportion of people the 3B algorithm classifies as having AA who have AA identified by chart review, defined as: Number of people identified with chart review to have AA and classified by the 3B algorithm to have AA / The number of people identified to have AA by the algorithm.

¥ Negative predictive value is the proportion of people the 3B algorithm classifies as not having AA identified by chart review, defined as: Number of people identified with chart review to not have AA and classified by the 3B algorithm to not have AA / The number of people identified to have AA by the algorithm.

€ This is the unadjusted predicted prevalence, defined as the percent of patients in the training sample predicted to be AA postiive using the corresponding risk score cut point. The unadjusted prevalence of AA positive patients in the training sample was 36.5% (511/1,400).

Source: Final report, page 57.

Reviewer Comment:
The 36-month simple ICD-9 code algorithm demonstrated a sensitivity of 0.39 and a PPV of 0.60 in the KPW validation data, both of which fall below the acceptable levels of performance established a priori.

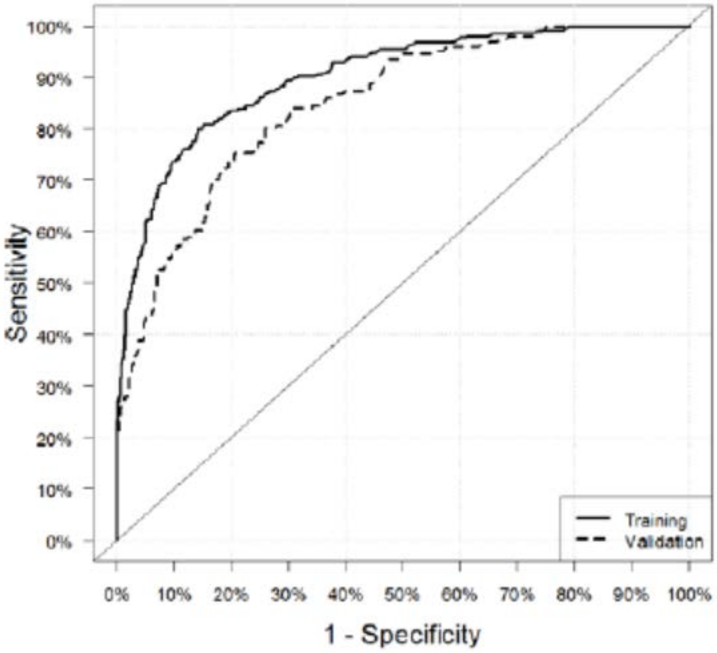
It is notable that the algorithm's specificity is relatively robust at the balancing point in both the training and validation data sets.

36-month EHR-enhanced algorithm

The EHR-enhanced algorithm did not perform appreciably better than its claims-only counterpart (based on the same set of candidate predictor variables but without the EHR-enhanced predictors). Its sensitivity and PPV of 0.590 and 0.566 in validation data (row 10 of Table 3.3.1.2.12) is not substantially different from that for the claims-only algorithm (0.582 and 0.572, respectively).

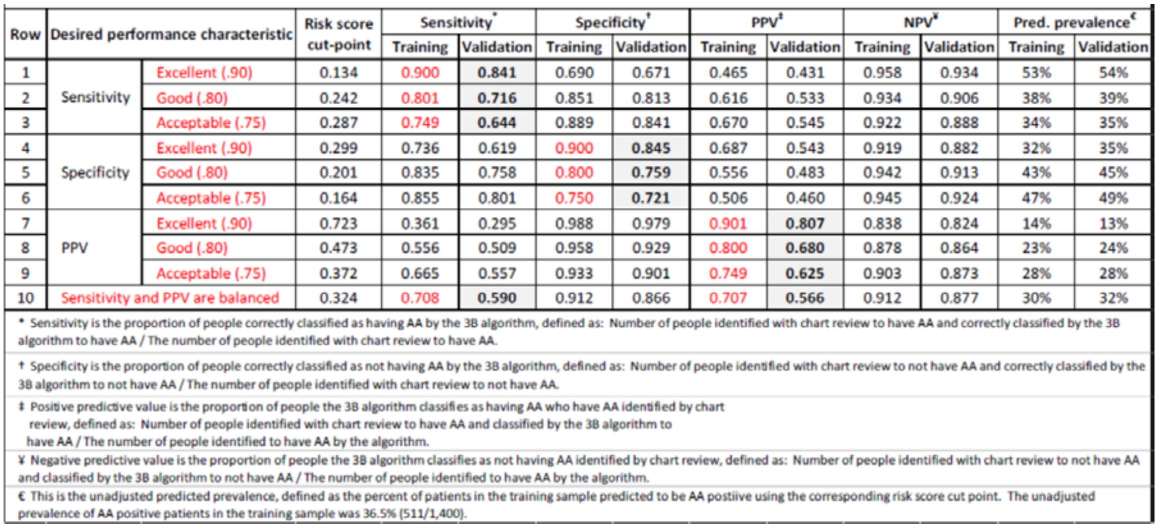
The study protocol acknowledged that the ambitious timeline of this study may limit the time and effort that could be devoted to implementing a truly best-case EHR-enhanced classification model. Time constraints turned out to be a significant factor in this work. Driven primarily by the study's focus on implementing and evaluating the claims-based classification models, these constraints prevented the study team from devoting a level of effort to development of the EHR-enhanced model that they considered sufficient to constitute it a strong algorithm development effort.

Figure 3.3.1.2.12. ROC curve for the 36-month EHR-enhanced algorithm.



Source: Final report, page 58.

Table 3.3.1.2.12. 36-month EHR-enhanced AA classification algorithm performance characteristics (sensitivity, specificity, PPV, NPV, predicted prevalence) in KPW training and validation samples for cut points of the AA risk score with desired performance characteristics selected based on training data.



Row	Desired performance characteristic		Risk score cut-point	Sensitivity*		Specificity†		PPV‡		NPV¥		Pred. prevalence€	
				Training	Validation	Training	Validation	Training	Validation	Training	Validation	Training	Validation
1	Sensitivity	Excellent (.90)	0.134	0.900	**0.841**	0.690	0.671	0.465	0.431	0.958	0.934	53%	54%
2		Good (.80)	0.242	0.801	**0.716**	0.851	0.813	0.616	0.533	0.934	0.906	38%	39%
3		Acceptable (.75)	0.287	0.749	**0.644**	0.889	0.841	0.670	0.545	0.922	0.888	34%	35%
4	Specificity	Excellent (.90)	0.299	0.736	0.619	0.900	**0.845**	0.687	0.543	0.919	0.882	32%	35%
5		Good (.80)	0.201	0.835	0.758	0.800	**0.759**	0.556	0.483	0.942	0.913	43%	45%
6		Acceptable (.75)	0.164	0.855	0.801	0.750	**0.721**	0.506	0.460	0.945	0.924	47%	49%
7	PPV	Excellent (.90)	0.723	0.361	0.295	0.988	0.979	0.901	**0.807**	0.838	0.824	14%	13%
8		Good (.80)	0.473	0.556	0.509	0.958	0.929	0.800	**0.680**	0.878	0.864	23%	24%
9		Acceptable (.75)	0.372	0.665	0.557	0.933	0.901	0.749	**0.625**	0.903	0.873	28%	28%
10	Sensitivity and PPV are balanced		0.324	0.708	**0.590**	0.912	0.866	0.707	**0.566**	0.912	0.877	30%	32%

* Sensitivity is the proportion of people correctly classified as having AA by the 3B algorithm, defined as: Number of people identified with chart review to have AA and correctly classified by the 3B algorithm to have AA / The number of people identified with chart review to have AA.

† Specificity is the proportion of people correctly classified as not having AA by the 3B algorithm, defined as: Number of people identified with chart review to not have AA and correctly classified by the 3B algorithm to not have AA / The number of people identified with chart review to not have AA.

‡ Positive predictive value is the proportion of people the 3B algorithm classifies as having AA who have AA identified by chart review, defined as: Number of people identified with chart review to have AA and classified by the 3B algorithm to have AA / The number of people identified to have AA by the algorithm.

¥ Negative predictive value is the proportion of people the 3B algorithm classifies as not having AA identified by chart review, defined as: Number of people identified with chart review to not have AA and classified by the 3B algorithm to not have AA / The number of people identified to have AA by the algorithm.

€ This is the unadjusted predicted prevalence, defined as the percent of patients in the training sample predicted to be AA postiive using the corresponding risk score cut point. The unadjusted prevalence of AA positive patients in the training sample was 36.5% (511/1,400).

Source: Final report, page 59.

Reviewer Comment:
The 36-month EHR-enhanced algorithm demonstrated a sensitivity of 0.59 and a PPV of 0.57 in the KPW validation data at the balancing point, both of which fall below the acceptable levels of performance established a priori.

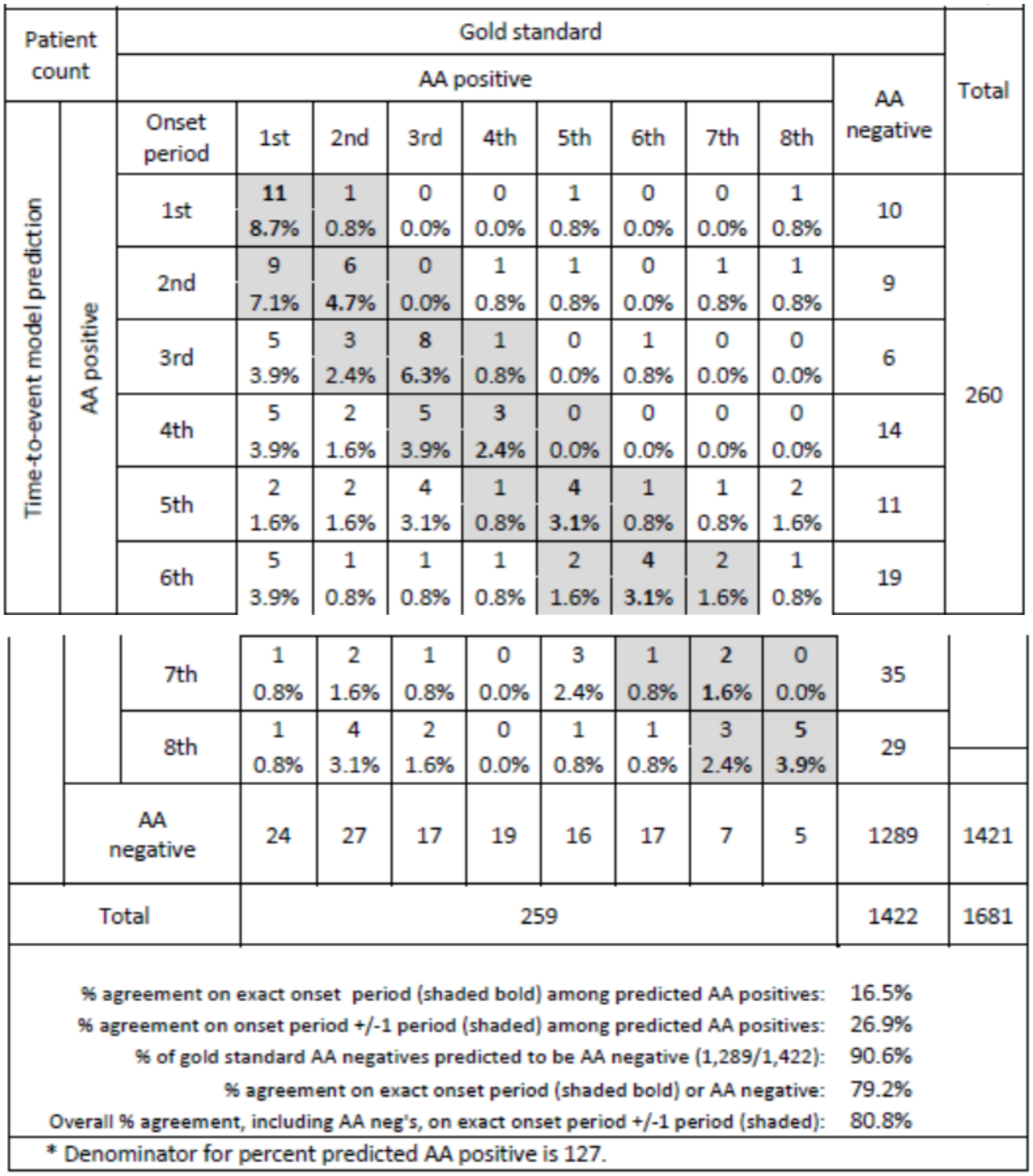
It is notable that the algorithm's specificity is relatively robust at the balancing point in both the training and validation data sets.

It is possible that limiting the population to ER/LA users resulted in a narrow study sample. Including long-term analgesic opioid users, regardless of formulation, may have provided more robust results.

Algorithm to predict AA onset

A total of 1,681 patients met eligibility criteria for the analysis to predict AA onset via the time-to-event model, including 259 (15.4%) determined to be incident AA positives during the 720-day follow up period. The time-to-event model generated a risk score for each patient and each 90-day time period. Using a cut point of 0.192 as the risk score threshold above which patients were predicted to experience AA onset, the model assigned one of eight onset periods to each of 260 patients with risk scores above the cut point. For comparison, there were 259 patients with AA onset during the same follow up periods according to the gold standard. Table 3.3.1.2.13 cross tabulates the model's predictions to the gold standard for all 1,681 patients in the analysis. Counts of patients along the diagonal (shaded, bold font) indicate exact agreement between the model and the gold standard with respect to onset period among patients predicted to be AA incident during follow up. Counts of patients in the shaded cells indicate agreement +/-1 period (i.e., the onset period predicted by the model is either the same period indicated by the gold standard or the period before or after). Also as shown in Table 3.3.1.2.13, among patients the model predicted to be AA positive there was exact agreement with the gold standard regarding period of AA onset for 16.5% of patients, corresponding to a kappa statistic of 0.264. Relaxing the agreement criterion to include the period before or after the period indicated by the gold standard increase the percentage agreement to 26.9%. The model and the gold standard agreed that AA was not present (AA negative) for 90.6% of patients (1,289/1,422). Accordingly, overall agreement between the model and the gold standard regarding the exact period of onset or that AA was not present was relatively high, at 79.2%. This overall agreement increased slightly, to 80.8%, if the criterion for agreement on AA positives is relaxed to +/-1 period.

Table 3.3.1.2.13. Counts of patients predicted to be abuse/addiction (AA) positive (and percent* of predicted AA positive) by the time-to-event model (rows) compared to the manually-abstracted gold standard for each of eight 90-day onset periods following each patient's study index date.



Patient count			Gold standard									Total
			AA positive								AA negative	
Time-to-event model prediction	AA positive	Onset period	1st	2nd	3rd	4th	5th	6th	7th	8th		
		1st	**11** **8.7%**	1 0.8%	0 0.0%	0 0.0%	1 0.8%	0 0.0%	0 0.0%	1 0.8%	10	260
		2nd	9 7.1%	**6** **4.7%**	0 0.0%	1 0.8%	1 0.8%	0 0.0%	1 0.8%	1 0.8%	9	
		3rd	5 3.9%	3 2.4%	**8** **6.3%**	1 0.8%	0 0.0%	1 0.8%	0 0.0%	0 0.0%	6	
		4th	5 3.9%	2 1.6%	5 3.9%	**3** **2.4%**	0 0.0%	0 0.0%	0 0.0%	0 0.0%	14	
		5th	2 1.6%	2 1.6%	4 3.1%	1 0.8%	**4** **3.1%**	1 0.8%	1 0.8%	2 1.6%	11	
		6th	5 3.9%	1 0.8%	1 0.8%	1 0.8%	2 1.6%	**4** **3.1%**	2 1.6%	1 0.8%	19	
		7th	1 0.8%	2 1.6%	1 0.8%	0 0.0%	3 2.4%	1 0.8%	**2** **1.6%**	0 0.0%	35	
		8th	1 0.8%	4 3.1%	2 1.6%	0 0.0%	1 0.8%	1 0.8%	3 2.4%	**5** **3.9%**	29	
	AA negative		24	27	17	19	16	17	7	5	1289	1421
Total			259								1422	1681

% agreement on exact onset period (shaded bold) among predicted AA positives: 16.5%

% agreement on onset period +/-1 period (shaded) among predicted AA positives: 26.9%

% of gold standard AA negatives predicted to be AA negative (1,289/1,422): 90.6%

% agreement on exact onset period (shaded bold) or AA negative: 79.2%

Overall % agreement, including AA neg's, on exact onset period +/-1 period (shaded): 80.8%

* Denominator for percent predicted AA positive is 127.

Source: Final report, pages 60-61.

Reviewer Comment:

These results suggest that definitively identifying AA onset is difficult even with manual chart review. The AA onset prediction algorithm was insufficient to accurately identify onset of AA. It is not clear that this secondary objective pertains to the goals of the PMR.

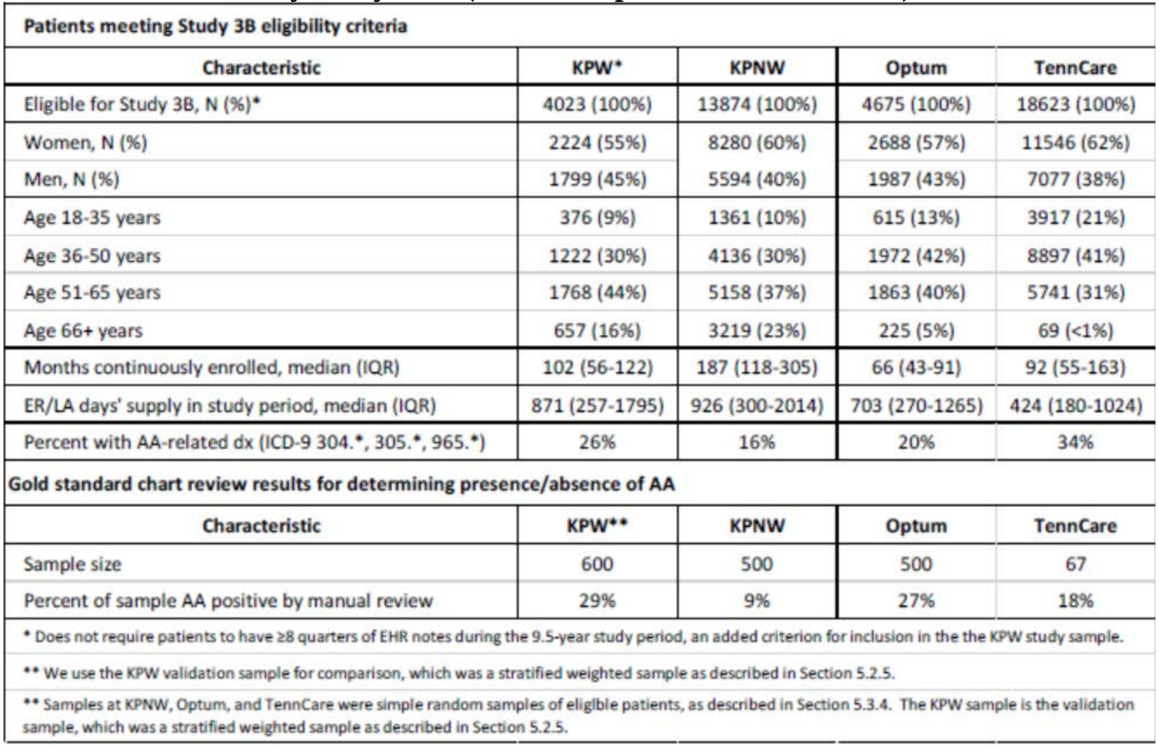
3.3.2 Secondary Study Sites

Results in the secondary study sites

Performance of the 36-month claims-based algorithm in secondary study sites is of reduced interest given its relatively poor performance in the primary study site. Nevertheless, investigators report here the planned analyses for this algorithm in each secondary site’s study sample.

The report first describes characteristics of the eligible populations at each secondary site, and compares them to the population from the primary study site. The report then presents results of applying the 36-month claims-based AA algorithm in each study site.

Table 3.3.2.1. Characteristics of sampled patients* meeting Study 3B eligibility criteria and results of the gold standard chart review for determining presence/absence of abuse/addiction (AA) at the primary study site (KPW) and each of the three secondary study sites (KPNW, Optum, and TennCare).**



Patients meeting Study 3B eligibility criteria

Characteristic	KPW*	KPNW	Optum	TennCare
Eligible for Study 3B, N (%)*	4023 (100%)	13874 (100%)	4675 (100%)	18623 (100%)
Women, N (%)	2224 (55%)	8280 (60%)	2688 (57%)	11546 (62%)
Men, N (%)	1799 (45%)	5594 (40%)	1987 (43%)	7077 (38%)
Age 18-35 years	376 (9%)	1361 (10%)	615 (13%)	3917 (21%)
Age 36-50 years	1222 (30%)	4136 (30%)	1972 (42%)	8897 (41%)
Age 51-65 years	1768 (44%)	5158 (37%)	1863 (40%)	5741 (31%)
Age 66+ years	657 (16%)	3219 (23%)	225 (5%)	69 (<1%)
Months continuously enrolled, median (IQR)	102 (56-122)	187 (118-305)	66 (43-91)	92 (55-163)
ER/LA days' supply in study period, median (IQR)	871 (257-1795)	926 (300-2014)	703 (270-1265)	424 (180-1024)
Percent with AA-related dx (ICD-9 304.*, 305.*, 965.*)	26%	16%	20%	34%

Gold standard chart review results for determining presence/absence of AA

Characteristic	KPW**	KPNW	Optum	TennCare
Sample size	600	500	500	67
Percent of sample AA positive by manual review	29%	9%	27%	18%

* Does not require patients to have ≥8 quarters of EHR notes during the 9.5-year study period, an added criterion for inclusion in the the KPW study sample.

** We use the KPW validation sample for comparison, which was a stratified weighted sample as described in Section 5.2.5.

** Samples at KPNW, Optum, and TennCare were simple random samples of eligible patients, as described in Section 5.3.4. The KPW sample is the validation sample, which was a stratified weighted sample as described in Section 5.2.5.

As shown in Table 3.3.2.1, the gender and age distributions were roughly comparable across study sites with some notable exceptions. The TennCare population had more eligible women (62%) than the other sites (ranging from 55% at KPW to 60% at KPNW). Age distributions at KPW and KPNW were similar to one another, though KPNW had more patients over 65 years of age. The Optum and TennCare eligible populations tended to be somewhat younger, with TennCare having the youngest population overall among the four study sites.

Reviewer Comment:

The low prevalence of AA positive charts in the KPNW sample relative to the other sites is notable. The KPNW individuals tended to be older, have longer periods of enrollment, and greater days' supply of ER/LA opioid analgesics than individuals at the other sites. This table underscores the small size of the TennCare sample relative to the other sites.

Duration of continuous enrollment periods varied considerably across the sites, with a median enrollment of 66 months at Optum (the shortest of the four), a median enrollment of 187 months at KPNW (the longest of the four), and KPW and TennCare medians of 102 and 92 months, respectively.

There is also considerable variation across the four study sites in total days' supply of ER/LA opioid analgesic medications dispensed to patients during each of their study-eligible periods of continuous enrollment, though these differences are highly correlated with differences in continuous enrollment, which is likely driving them. For example, median total days of prescribed ER/LA opioid analgesics for TennCare patients was 424 compared to 926 at KPNW, which is roughly proportional to the differences in duration of enrollment at these two sites.

The percentage of patients with any AA-related ICD-9 diagnosis codes in their medical records varied considerably across sites. The rate of 34% at TennCare is over double the rate of 16% observed at KPNW. The rate for Optum patients (20%) and KPW (26%) are in between.

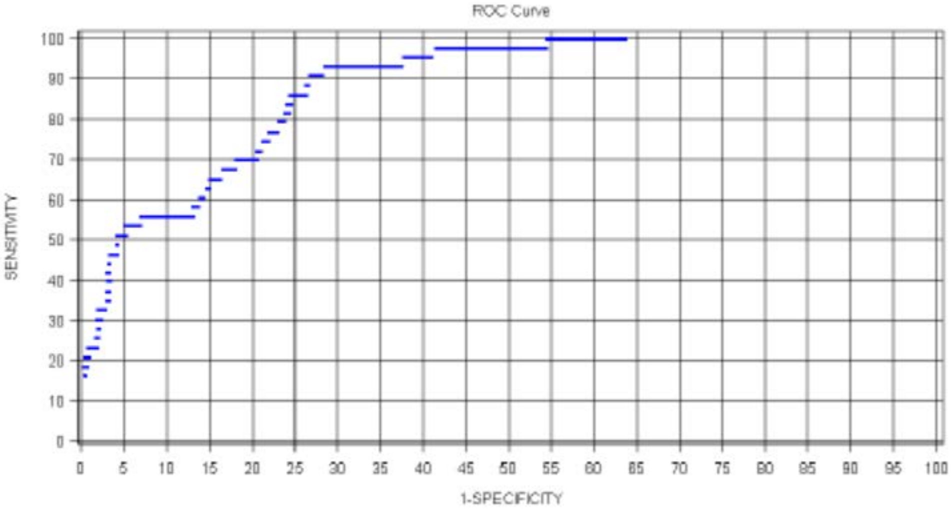
The percentages of patients determined to be AA positive by manual review also varied considerably across sites, from a low of 9% at KPNW to a high of 29% at KPW. This is somewhat surprising, given that these two healthcare systems are considered to be the most similar in terms of patient populations and care delivery practices among the four sites. The percentages of AA positives at Optum was similar to that at KPW (27% and 29%, respectively), and the rate at TennCare (18%) was in the middle of the distribution. However, differences between the manual records review data and processes at KPW and Optum, and questions about the generalizability of the TennCare sample to the larger Tennessee Medicaid population, warrant restraint in reading too much into these similarities and differences.

KPNW

Results of applying the 36-month claims-based version of the AA algorithm in the KPNW portability sample are summarized in Figure 3.3.2.2 and Table 3.3.2.2. The expected decline in performance when applying an algorithm developed at one site in another site is evident in row 10 of Table 3.3.2.2, which shows the model achieved sensitivity of 0.512 and PPV of 0.537 (compared to values of 0.582 and 0.572 in validation data at the primary study site).

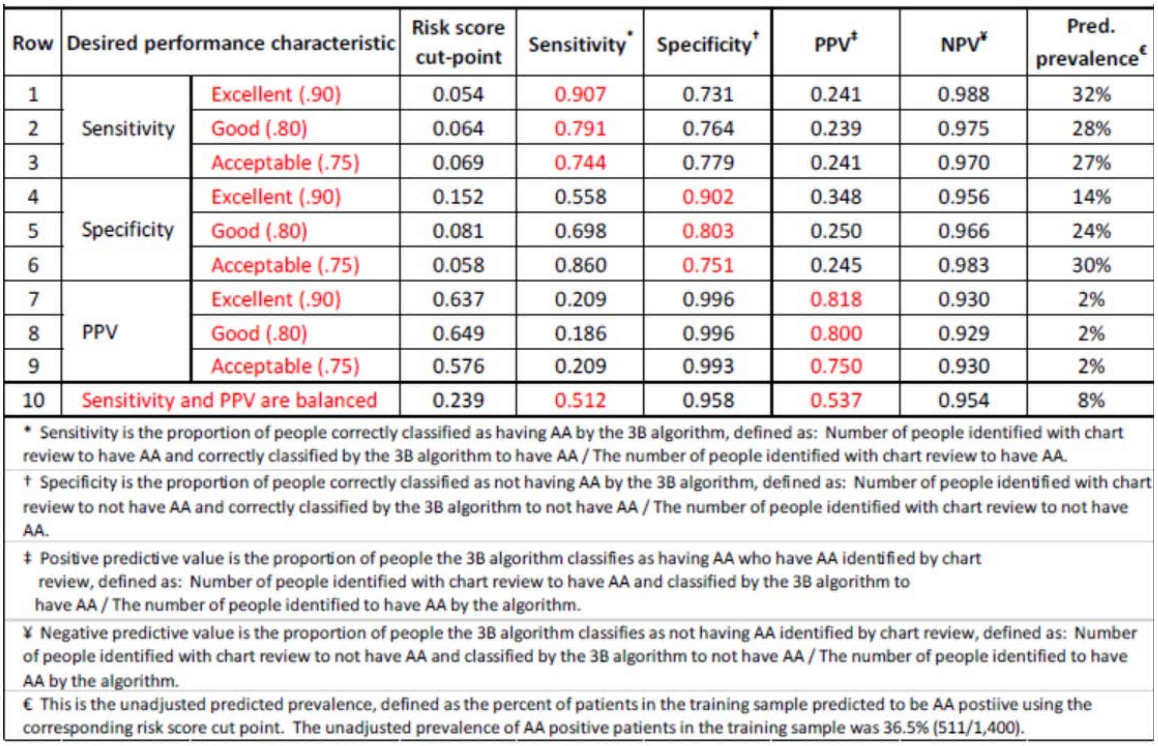
Also notable is the large difference in predicted prevalence in the KPNW data between a version of the algorithm that achieves high sensitivity (yielding a predicted prevalence of 32% with sensitivity at 0.907, Table 3.3.2.2 row 1) and a version that maximizes PPV (yielding a predicted prevalence of 2% with PPV at 0.818, the highest PPV possible in KPNW data, Table 3.3.2.2 row 7), a fifteen-fold difference. The comparable difference in predicted prevalence in the KPW data is just over a four-fold difference.

Figure 3.3.2.2. ROC curve for the 36-month AA algorithm in KPNW data.



Source: Final report, page 64.

Table 3.3.2.2. AA classification algorithm performance characteristics (sensitivity, specificity, PPV, NPV, predicted prevalence) in the sample of 500 KPNW subjects for selected risk score cut points based on KPW training data.



Row	Desired performance characteristic		Risk score cut-point	Sensitivity*	Specificity†	PPV‡	NPV¥	Pred. prevalence€
1	Sensitivity	Excellent (.90)	0.054	0.907	0.731	0.241	0.988	32%
2		Good (.80)	0.064	0.791	0.764	0.239	0.975	28%
3		Acceptable (.75)	0.069	0.744	0.779	0.241	0.970	27%
4	Specificity	Excellent (.90)	0.152	0.558	0.902	0.348	0.956	14%
5		Good (.80)	0.081	0.698	0.803	0.250	0.966	24%
6		Acceptable (.75)	0.058	0.860	0.751	0.245	0.983	30%
7	PPV	Excellent (.90)	0.637	0.209	0.996	0.818	0.930	2%
8		Good (.80)	0.649	0.186	0.996	0.800	0.929	2%
9		Acceptable (.75)	0.576	0.209	0.993	0.750	0.930	2%
10	Sensitivity and PPV are balanced		0.239	0.512	0.958	0.537	0.954	8%

* Sensitivity is the proportion of people correctly classified as having AA by the 3B algorithm, defined as: Number of people identified with chart review to have AA and correctly classified by the 3B algorithm to have AA / The number of people identified with chart review to have AA.

† Specificity is the proportion of people correctly classified as not having AA by the 3B algorithm, defined as: Number of people identified with chart review to not have AA and correctly classified by the 3B algorithm to not have AA / The number of people identified with chart review to not have AA.

‡ Positive predictive value is the proportion of people the 3B algorithm classifies as having AA who have AA identified by chart review, defined as: Number of people identified with chart review to have AA and classified by the 3B algorithm to have AA / The number of people identified to have AA by the algorithm.

¥ Negative predictive value is the proportion of people the 3B algorithm classifies as not having AA identified by chart review, defined as: Number of people identified with chart review to not have AA and classified by the 3B algorithm to not have AA / The number of people identified to have AA by the algorithm.

€ This is the unadjusted predicted prevalence, defined as the percent of patients in the training sample predicted to be AA postiive using the corresponding risk score cut point. The unadjusted prevalence of AA positive patients in the training sample was 36.5% (511/1,400).

Source: Final report, page 65.

Reviewer Comment:
The 36-month algorithm demonstrated a sensitivity of 0.51 and a PPV of 0.54 in the KPNW sample at the balancing point, both of which fall below the acceptable levels of performance established a priori.

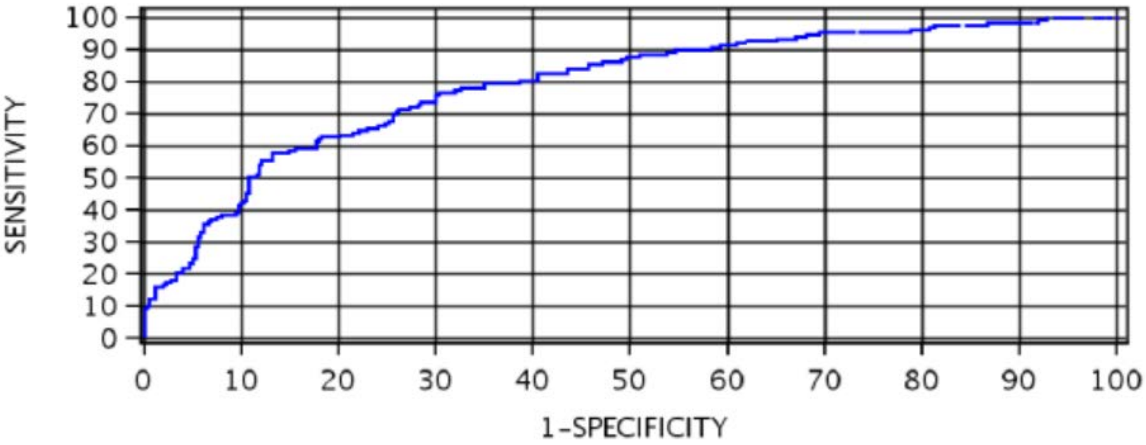
It is notable that the algorithm's specificity is relatively robust at the balancing point in both the training and validation data sets.

Optum

Results of applying the 36-month claims-based version of the AA algorithm in the Optum portability sample are summarized in Figure 3.3.2.3 and Table 3.3.2.3. The algorithm's performance in Optum data did not decline compared to its performance in the KPW data. The results are similar, with slightly better performance in the Optum data than the KPW validation sample. This is evident, for example, when comparing algorithm performance using cut points of the risk score that balance sensitivity and PPV. Sensitivity and PPV in the Optum data at this cut point are both 0.59, which are slightly higher than the corresponding values observed in KPW validation data (0.58 and 0.57, respectively).

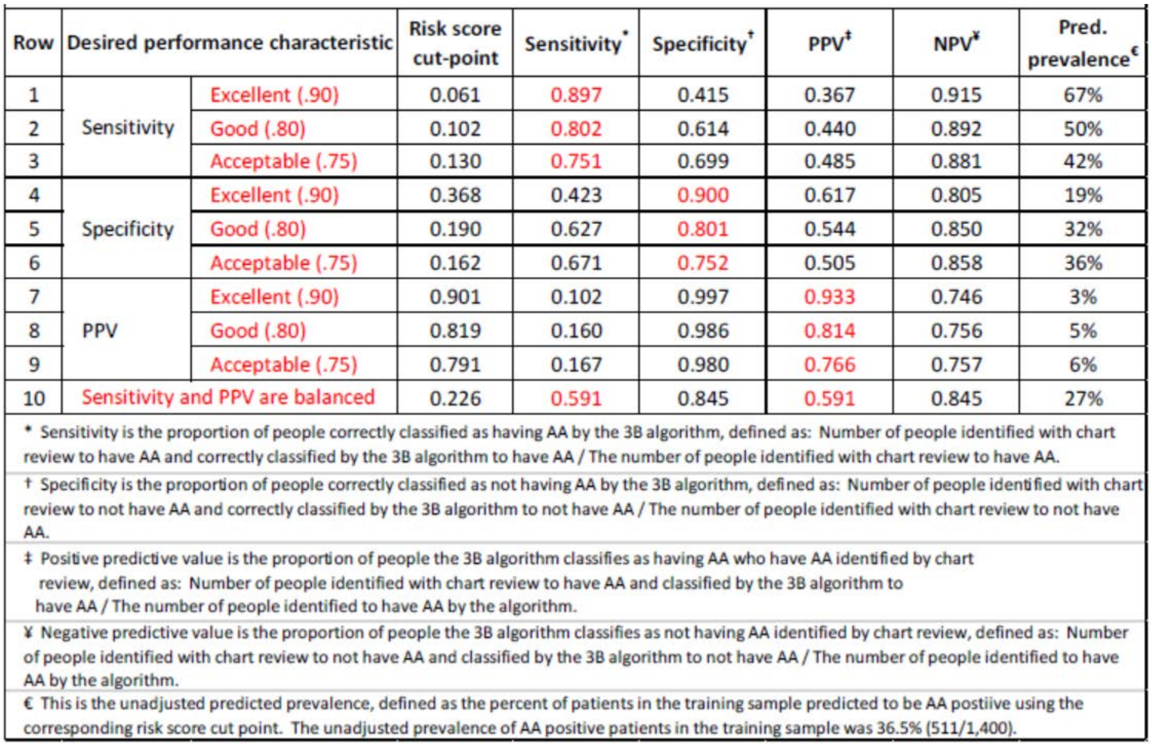
Despite this slight improvement, the algorithm lacks performance characteristics that would justify using it to identify and investigate risk factors associated with AA positive and AA negative patients. Investigators also see a fifteen-fold difference in predicted prevalence values in the Optum data when comparing versions of the algorithm that maximize sensitivity versus versions that maximize PPV (Table 3.3.2.2, far right column).

Figure 3.3.2.3. ROC curve for the 36-month AA algorithm in Optum data.



Source: Final report, page 66.

Table 3.3.2.3. AA classification algorithm performance characteristics (sensitivity, specificity, PPV, NPV, predicted prevalence) in the sample of 500 Optum subjects for selected risk score cut points based on KPW training data.



Row	Desired performance characteristic		Risk score cut-point	Sensitivity*	Specificity†	PPV‡	NPV¥	Pred. prevalence€
1	Sensitivity	Excellent (.90)	0.061	0.897	0.415	0.367	0.915	67%
2		Good (.80)	0.102	0.802	0.614	0.440	0.892	50%
3		Acceptable (.75)	0.130	0.751	0.699	0.485	0.881	42%
4	Specificity	Excellent (.90)	0.368	0.423	0.900	0.617	0.805	19%
5		Good (.80)	0.190	0.627	0.801	0.544	0.850	32%
6		Acceptable (.75)	0.162	0.671	0.752	0.505	0.858	36%
7	PPV	Excellent (.90)	0.901	0.102	0.997	0.933	0.746	3%
8		Good (.80)	0.819	0.160	0.986	0.814	0.756	5%
9		Acceptable (.75)	0.791	0.167	0.980	0.766	0.757	6%
10	Sensitivity and PPV are balanced		0.226	0.591	0.845	0.591	0.845	27%

* Sensitivity is the proportion of people correctly classified as having AA by the 3B algorithm, defined as: Number of people identified with chart review to have AA and correctly classified by the 3B algorithm to have AA / The number of people identified with chart review to have AA.

† Specificity is the proportion of people correctly classified as not having AA by the 3B algorithm, defined as: Number of people identified with chart review to not have AA and correctly classified by the 3B algorithm to not have AA / The number of people identified with chart review to not have AA.

‡ Positive predictive value is the proportion of people the 3B algorithm classifies as having AA who have AA identified by chart review, defined as: Number of people identified with chart review to have AA and classified by the 3B algorithm to have AA / The number of people identified to have AA by the algorithm.

¥ Negative predictive value is the proportion of people the 3B algorithm classifies as not having AA identified by chart review, defined as: Number of people identified with chart review to not have AA and classified by the 3B algorithm to not have AA / The number of people identified to have AA by the algorithm.

€ This is the unadjusted predicted prevalence, defined as the percent of patients in the training sample predicted to be AA postiive using the corresponding risk score cut point. The unadjusted prevalence of AA positive patients in the training sample was 36.5% (511/1,400).

Source: Final report, page 67.

Reviewer Comment:

The 36-month algorithm demonstrated a sensitivity of 0.59 and a PPV of 0.59 in the Optum sample at the balancing point, both of which fall below the acceptable levels of performance established a priori.

It is notable that the algorithm's specificity is relatively robust at the balancing point in both the training and validation data sets.

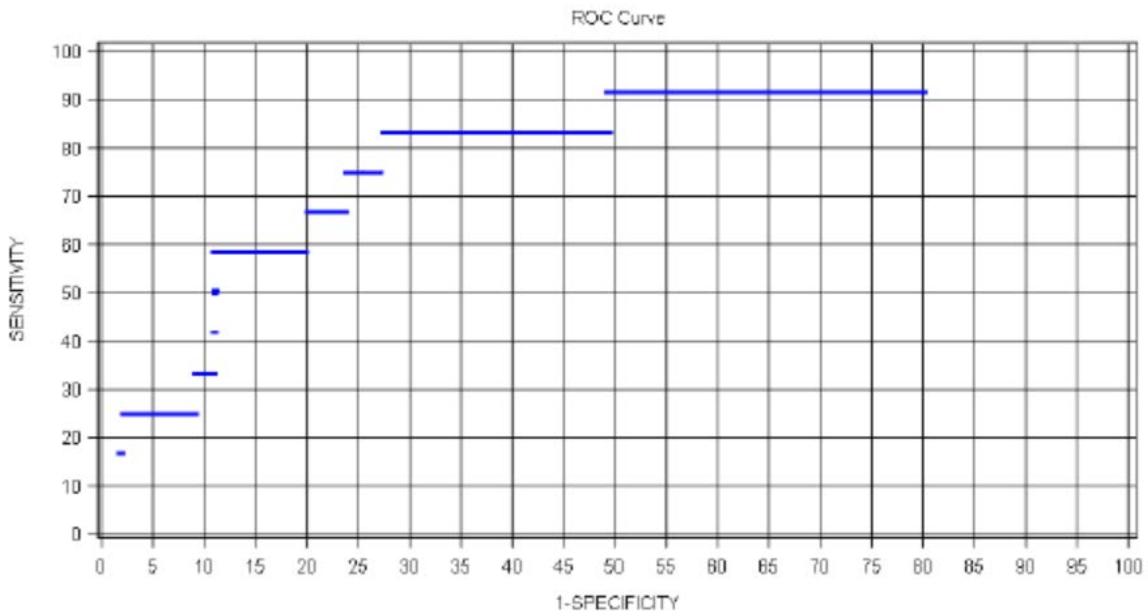
TennCare

Results of applying the 36-month claims-based version of the AA algorithm in the TennCare portability sample of 67 patients receiving care in the Vanderbilt University Clinic are summarized in Figure 3.3.2.4 and Table 3.3.2.4. The expected decline in performance when applying an algorithm developed at one site in another site is evident in row 10 of Table 3.3.2.4, which shows the model achieved sensitivity of 0.58 and PPV of 0.54, which is lower than the corresponding performance metrics observed in the KPW validation sample, which were 0.58 and 0.57, respectively. Notably, there was no risk score cut point that yielded acceptable PPV in TennCare data. As shown in row 9 of Table 3.3.2.4, the highest PPV attainable by this algorithm when applied to the TennCare sample was 0.54. Depending on the risk score cut point chosen, predicted prevalence in the TennCare data ranged from a low of 19% to a high of 57%. This approximately three-fold difference in predicted prevalence is comparable to that observed in the primary study site, but it is based on a small sample (N=67) and as already noted, there was no cut point that provided high PPV; had there been such a cut point, the range in predicted prevalence would likely have been larger.

As was the case at the other two secondary study sites, when applied to TennCare data the algorithm lacked performance characteristics that would justify using it to identify and investigate risk factors associated with AA positive and AA negative patients.

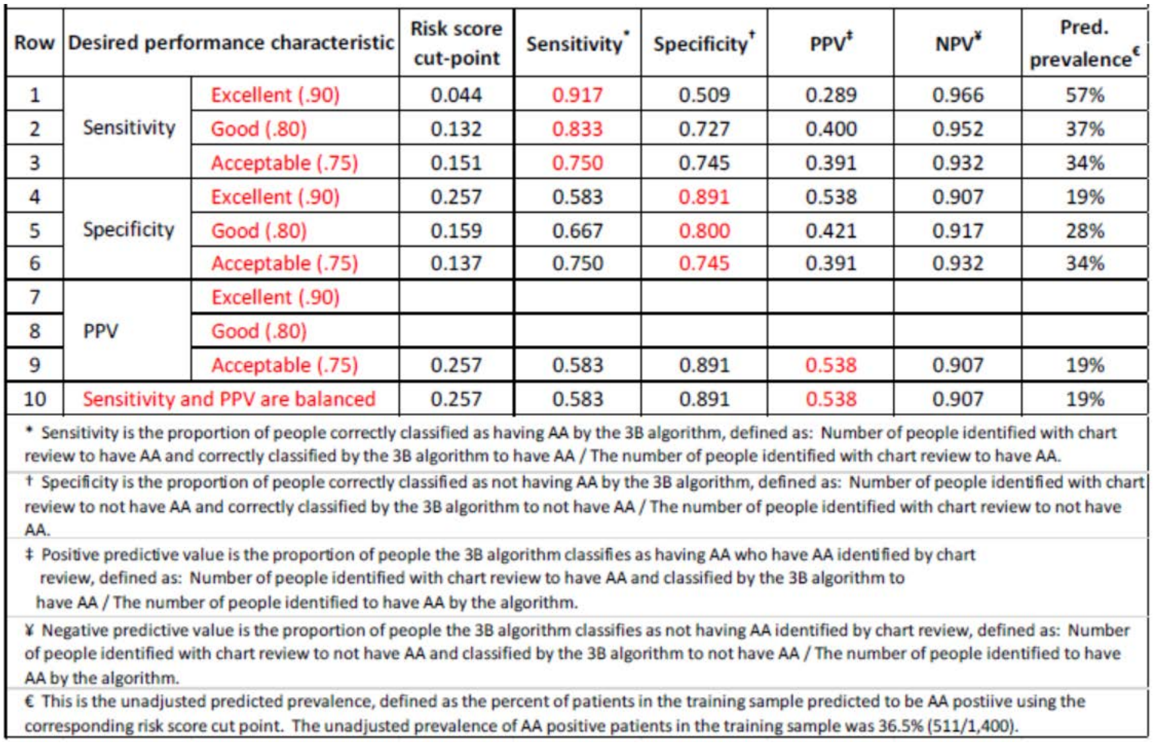
The results in TennCare data should be interpreted cautiously because of the small sample size and concerns about the generalizability of the sample selected from patients receiving the majority of their care from the Vanderbilt University Clinic, which may not be representative of typical Tennessee Medicaid enrollees.

Figure 3.3.2.4. ROC curve for the 36-month AA algorithm in TennCare data.



Source: Final report, page 68.

Table 3.3.2.4. AA classification algorithm performance characteristics (sensitivity, specificity, PPV, NPV, predicted prevalence) in the sample of 67 TennCare subjects for selected risk score cut points based on KPW training data.



Row	Desired performance characteristic		Risk score cut-point	Sensitivity*	Specificity†	PPV‡	NPV¥	Pred. prevalence€
1	Sensitivity	Excellent (.90)	0.044	0.917	0.509	0.289	0.966	57%
2		Good (.80)	0.132	0.833	0.727	0.400	0.952	37%
3		Acceptable (.75)	0.151	0.750	0.745	0.391	0.932	34%
4	Specificity	Excellent (.90)	0.257	0.583	0.891	0.538	0.907	19%
5		Good (.80)	0.159	0.667	0.800	0.421	0.917	28%
6		Acceptable (.75)	0.137	0.750	0.745	0.391	0.932	34%
7	PPV	Excellent (.90)						
8		Good (.80)						
9		Acceptable (.75)	0.257	0.583	0.891	0.538	0.907	19%
10	Sensitivity and PPV are balanced		0.257	0.583	0.891	0.538	0.907	19%

* Sensitivity is the proportion of people correctly classified as having AA by the 3B algorithm, defined as: Number of people identified with chart review to have AA and correctly classified by the 3B algorithm to have AA / The number of people identified with chart review to have AA.

† Specificity is the proportion of people correctly classified as not having AA by the 3B algorithm, defined as: Number of people identified with chart review to not have AA and correctly classified by the 3B algorithm to not have AA / The number of people identified with chart review to not have AA.

‡ Positive predictive value is the proportion of people the 3B algorithm classifies as having AA who have AA identified by chart review, defined as: Number of people identified with chart review to have AA and classified by the 3B algorithm to have AA / The number of people identified to have AA by the algorithm.

¥ Negative predictive value is the proportion of people the 3B algorithm classifies as not having AA identified by chart review, defined as: Number of people identified with chart review to not have AA and classified by the 3B algorithm to not have AA / The number of people identified to have AA by the algorithm.

€ This is the unadjusted predicted prevalence, defined as the percent of patients in the training sample predicted to be AA postiive using the corresponding risk score cut point. The unadjusted prevalence of AA positive patients in the training sample was 36.5% (511/1,400).

Source: Final report, page 69.

Reviewer Comment:

The 36-month algorithm demonstrated a sensitivity of 0.58 and a PPV of 0.54 in the TennCare sample at the balancing point, both of which fall below the acceptable levels of performance established a priori.

It is notable that the algorithm's specificity is relatively robust at the balancing point in both the training and validation data sets.

The algorithm was unable to demonstrate acceptable sensitivity and PPV metrics at any of the primary or secondary sites.

4 DISCUSSION

The PMR letter to the OPC stated that study 3033-7 should be:

> *"An observational study to develop and validate an algorithm using coded medical terminologies to identify patients experiencing prescription opioid abuse or addiction, among patients receiving an ER/LA opioid analgesic."*

The algorithms demonstrated generally high levels of specificity across all study sites. However, **this study did not show that an automated algorithm based on medical claims data could determine which patient charts contained evidence of AA with adequate sensitivity and PPV to reliably inform regulatory safety concerns.** The findings of this study suggest that AA is a phenomenon that is not well-reflected in medical claims data. Supplementing the algorithm with NLP- and EHR-generated data did not markedly improve performance.

The study confirmed that the primary outcome of interest, identification of AA through chart review, was feasible with high reliability and high inter-rater reliability. The investigators' secondary outcome of interest – identifying AA onset date – was not readily feasible, even using manual chart abstraction.

The AA onset analysis – while of uncertain value in the context of this PMR – was notable in that many of the studied patients appeared to demonstrate AA before ER/LA exposure. As the authors note: "Over half of the study cohort had received >60 days' supply of SA opioids in the 6 months preceding their study index dates. Not surprisingly, almost half of the cohort determined to be AA positive by manual chart review had AA onset dates that preceded or coincided with the study index date." This finding suggests that the transition of individuals with AA to ER/LA opioid analgesics may be of an equal or greater concern than the transition of individuals dispensed ER/LA opioid analgesics to AA.
One possibility is that clinicians may be "channeling" patients exhibiting AA behaviors from immediate-release opioid analgesics to ER/LA opioid analgesics. Some prescribers may view ER/LA opioid analgesics' steady release of active pharmaceutical ingredient (API) as beneficial to avoid the reinforcing peak-and-trough psychotropic effects observed with immediate-release opioid analgesics. However, this presumed

pharmacokinetic benefit is undercut by the fact that ER/LA opioid analgesics typically contain more API than immediate-release opioid analgesics: patients demonstrating AA behaviors who are channeled to ER/LA opioid analgesics on this basis therefore may experience a reinforcement of the AA behavior based on the increased total API exposure. FDA has received anecdotal reports of this channeling phenomenon, and these results indicate that there may be value in performing additional research on the complex temporal relationship between ER/LA opioid analgesic exposure, AA behavior, and possible prescriber channeling. The observation of this phenomenon among patients enrolled in a Kaiser system – generally regarded as an upper-tier health insurance payor with stringent opioid analgesic prescribing policies – indicates that this phenomenon could certainly happen in other health care systems as well.

The secondary study sites represented different healthcare settings and patient populations (fee-for service commercial insurance [Optum], staff-model managed care [KPNW], and Medicaid [TennCare]). However, time and logistical constraints resulted in sample size limitations. As noted by the authors, "This limitation was inconsequential in the present study because the AA algorithm failed to perform well in the primary study site (KPW), where sufficient training and validation data were available. However, future studies intending to develop algorithms that perform well in multiple and diverse study settings should consider provisions for improving both the quantity and quality of data from secondary study sites."

These results raise questions about the appropriateness of selecting the study population to answer the question of interest. One is whether the KPW validation/gold standard population was somewhat different than even the general Kaiser patient population. In an integrated care system such as KP that has many restrictions on opioid analgesic use, the use of high-risk medications such as ER/LA opioid analgesics suggests that the patients receiving these drugs are unique in some way to have been granted an exception to overcome the administrative controls on opioid analgesic dispensing. One also wonders if the algorithm would have performed just as poorly if eligible individuals were those on long-term therapy, regardless of formulation (IR or ER/LA). Finally, it would have been helpful for the investigators to have clarified the analytic approach to "balancing" sensitivity and PPV for the purposes of identifying a cut-off risk score, as this was a central consideration in the algorithm's performance.

The study's results may be a function of the phenomenon under investigation rather than a flawed study design. AA is a complex, chronic, and covert phenomenon. Definitive diagnoses of the condition are difficult to ascertain from EHR charts, even when reviewed by trained abstractors. Importantly, although the algorithm was not successful, other findings in this study could generate hypotheses to be tested about prescribing practices and patient characteristics that could support future regulatory decision-making. For example, many patients had evidence of AA prior to ER/LA opioid analgesic exposure. This suggests potential value in performing additional research on the complex temporal relationship between ER/LA opioid analgesic exposure, AA behavior, and possible prescriber channeling. Additionally, data on prescriber characteristics and perceptions of patient risk could inform future regulatory discussions surrounding the safe prescribing and use of ER/LA opioid analgesics.

5 CONCLUSION

5.1 Sponsor Conclusions

> *"Study 3B successfully developed high-quality gold standard data for a large, 2,000 patient sample, based on manual review of patient EHR charts. These data provided ample opportunity to assess the feasibility of developing an automated algorithm for identifying AA based on widely available structured claims data.*
>
> *Despite considerable effort and consideration of a very large number of potential predictors of AA* ***Study 3B did not yield a high-performing automated algorithm for identifying AA based on widely available structured claims data. Nor did the study results yield encouragement that development of such an algorithm is feasible.***
>
> *Investigators recommend that the AA algorithm developed in Study 3B not be used in Study 1B, even as an exploratory aim. The rates of misclassification of AA positive and AA negative patients observed in Study 3B indicate that attempting to use its algorithm-generated classifications regarding the presence or absence of AA as the basis for studying risk factors associated with these outcomes may be as likely to mislead as it would be to inform."*

5.2 Reviewer Conclusions

The algorithms demonstrated generally high levels of specificity across all study sites. However, **the algorithm had poor performance at discriminating between patients experiencing AA and those not experiencing AA in both the primary and secondary study sites.** The risk score cut-off points balancing/optimizing sensitivity and positive predictive value resulted in algorithms lacking **adequate sensitivity and PPV to reliably inform regulatory safety concerns**. In many analyses, the algorithm performed little better than chance with respect to these metrics. It would have been helpful for the investigators to have clarified the analytic approach to "balancing" sensitivity and PPV for the purposes of identifying a cut-off risk score, as this was a central consideration in the algorithm's performance.

Compared to the gold standard of chart review, the developed claims-based algorithm was not able to identify the presence of opioid addiction or abuse with a sufficient blend of sensitivity and positive predictive value to warrant further use. The AA onset algorithm was similarly inadequate, though the results indicate that there may be value in performing additional research on the complex temporal relationship between ER/LA opioid analgesic exposure, AA behavior, and possible prescriber channeling.

Despite the algorithm's poor performance, it appears to the reviewer that – based on the study's methodology and analytic approaches – adequate effort was made to develop an algorithm that successfully identified patients with AA using claims data. The postmarketing requirement should be considered fulfilled based on this submission.

6 RECOMMENDATIONS

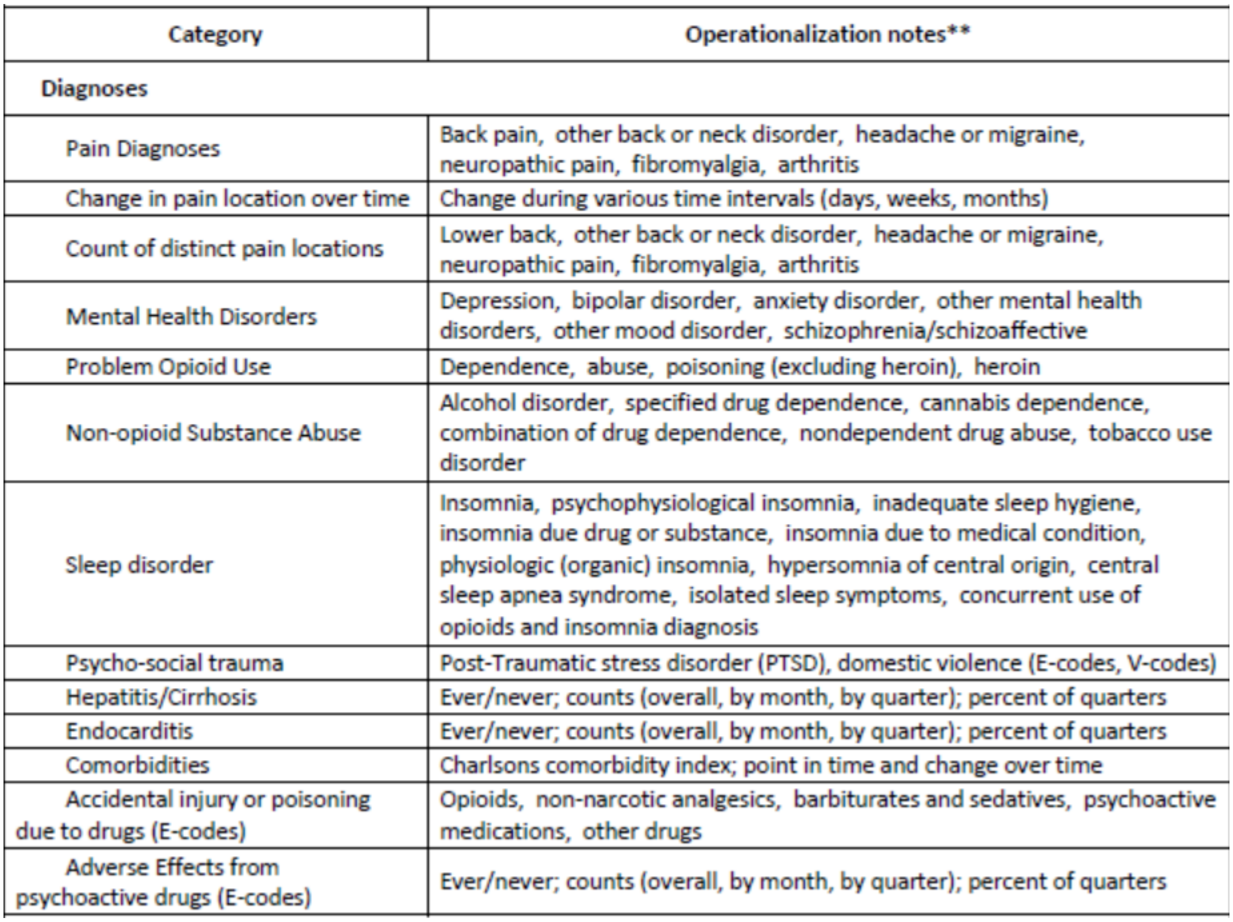
The algorithms developed in this study are inadequate to reliably inform regulatory safety concerns and should not be used in any additional studies involving the ER/LA PMR studies.

The postmarketing requirement should be considered fulfilled based on this submission, and the authors should strongly consider publishing these results so that other investigators can build on this work.

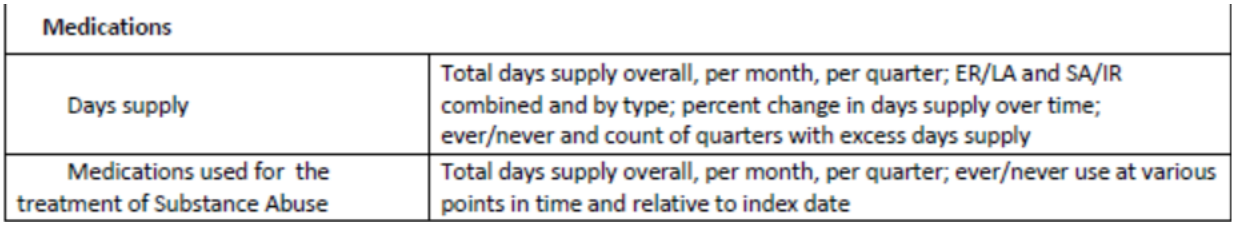
7 APPENDICES

7.1 APPENDIX A: CATEGORIES OF POTENTIAL PREDICTOR VARIABLES OPERATIONALIZED FROM MEDICAL CLAIMS DATA.

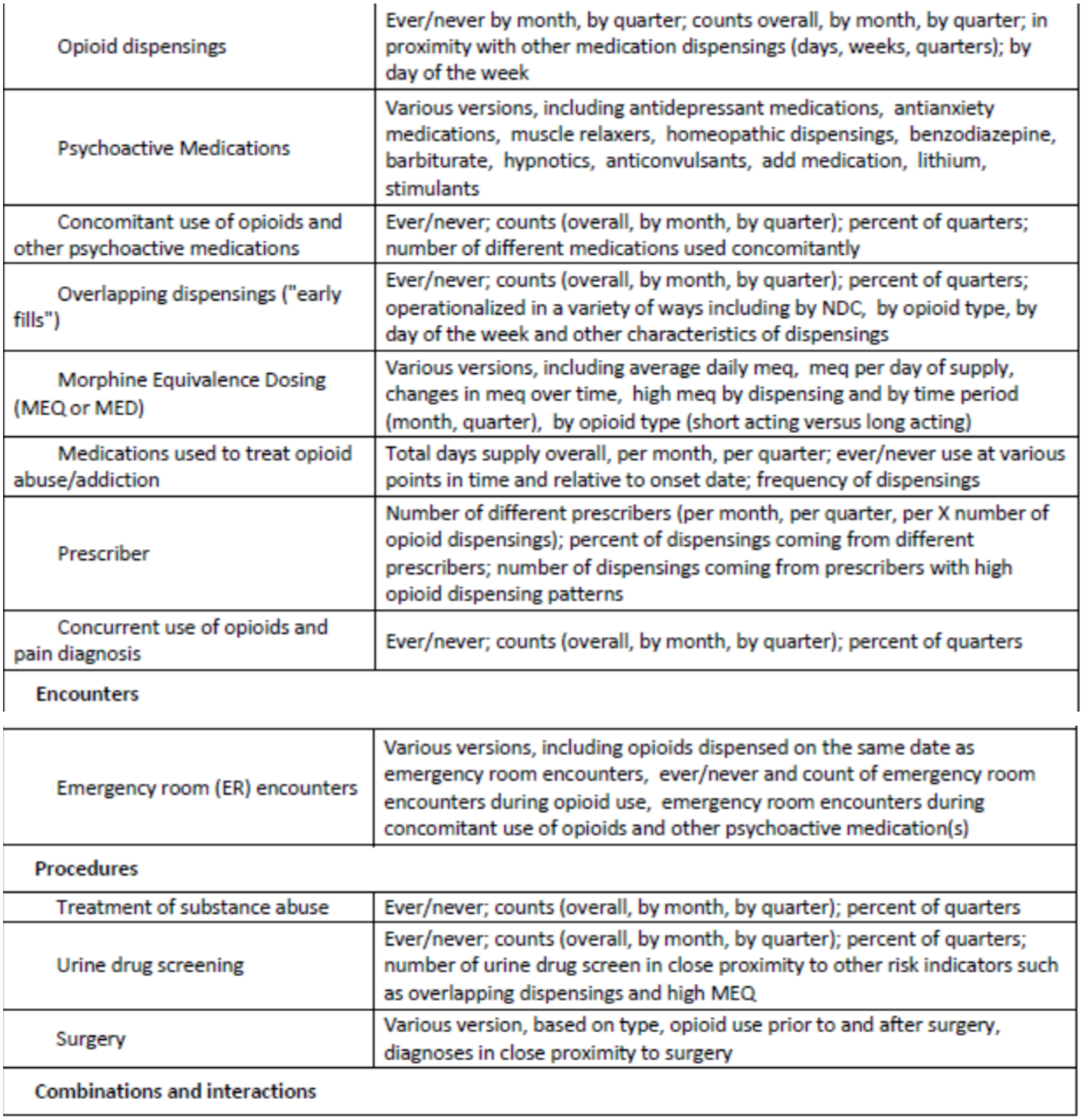
Variables include diagnoses, medications, encounters, procedures and other combinations or interactions considered for inclusion in the classification algorithm to identify patients with opioid-related abuse/addiction (AA)* by category with operationalization notes.



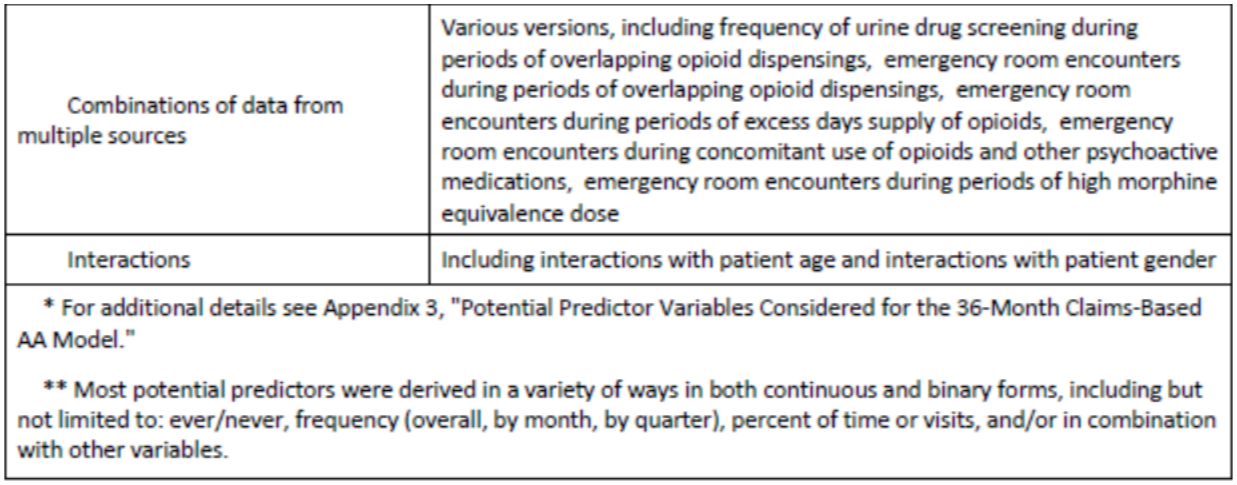
Category	Operationalization notes**
Diagnoses	
Pain Diagnoses	Back pain, other back or neck disorder, headache or migraine, neuropathic pain, fibromyalgia, arthritis
Change in pain location over time	Change during various time intervals (days, weeks, months)
Count of distinct pain locations	Lower back, other back or neck disorder, headache or migraine, neuropathic pain, fibromyalgia, arthritis
Mental Health Disorders	Depression, bipolar disorder, anxiety disorder, other mental health disorders, other mood disorder, schizophrenia/schizoaffective
Problem Opioid Use	Dependence, abuse, poisoning (excluding heroin), heroin
Non-opioid Substance Abuse	Alcohol disorder, specified drug dependence, cannabis dependence, combination of drug dependence, nondependent drug abuse, tobacco use disorder
Sleep disorder	Insomnia, psychophysiological insomnia, inadequate sleep hygiene, insomnia due drug or substance, insomnia due to medical condition, physiologic (organic) insomnia, hypersomnia of central origin, central sleep apnea syndrome, isolated sleep symptoms, concurrent use of opioids and insomnia diagnosis
Psycho-social trauma	Post-Traumatic stress disorder (PTSD), domestic violence (E-codes, V-codes)
Hepatitis/Cirrhosis	Ever/never; counts (overall, by month, by quarter); percent of quarters
Endocarditis	Ever/never; counts (overall, by month, by quarter); percent of quarters
Comorbidities	Charlsons comorbidity index; point in time and change over time
Accidental injury or poisoning due to drugs (E-codes)	Opioids, non-narcotic analgesics, barbiturates and sedatives, psychoactive medications, other drugs
Adverse Effects from psychoactive drugs (E-codes)	Ever/never; counts (overall, by month, by quarter); percent of quarters



Medications	
Days supply	Total days supply overall, per month, per quarter; ER/LA and SA/IR combined and by type; percent change in days supply over time; ever/never and count of quarters with excess days supply
Medications used for the treatment of Substance Abuse	Total days supply overall, per month, per quarter; ever/never use at various points in time and relative to index date



Opioid dispensings	Ever/never by month, by quarter; counts overall, by month, by quarter; in proximity with other medication dispensings (days, weeks, quarters); by day of the week
Psychoactive Medications	Various versions, including antidepressant medications, antianxiety medications, muscle relaxers, homeopathic dispensings, benzodiazepine, barbiturate, hypnotics, anticonvulsants, add medication, lithium, stimulants
Concomitant use of opioids and other psychoactive medications	Ever/never; counts (overall, by month, by quarter); percent of quarters; number of different medications used concomitantly
Overlapping dispensings ("early fills")	Ever/never; counts (overall, by month, by quarter); percent of quarters; operationalized in a variety of ways including by NDC, by opioid type, by day of the week and other characteristics of dispensings
Morphine Equivalence Dosing (MEQ or MED)	Various versions, including average daily meq, meq per day of supply, changes in meq over time, high meq by dispensing and by time period (month, quarter), by opioid type (short acting versus long acting)
Medications used to treat opioid abuse/addiction	Total days supply overall, per month, per quarter; ever/never use at various points in time and relative to onset date; frequency of dispensings
Prescriber	Number of different prescribers (per month, per quarter, per X number of opioid dispensings); percent of dispensings coming from different prescribers; number of dispensings coming from prescribers with high opioid dispensing patterns
Concurrent use of opioids and pain diagnosis	Ever/never; counts (overall, by month, by quarter); percent of quarters
Encounters	
Emergency room (ER) encounters	Various versions, including opioids dispensed on the same date as emergency room encounters, ever/never and count of emergency room encounters during opioid use, emergency room encounters during concomitant use of opioids and other psychoactive medication(s)
Procedures	
Treatment of substance abuse	Ever/never; counts (overall, by month, by quarter); percent of quarters
Urine drug screening	Ever/never; counts (overall, by month, by quarter); percent of quarters; number of urine drug screen in close proximity to other risk indicators such as overlapping dispensings and high MEQ
Surgery	Various version, based on type, opioid use prior to and after surgery, diagnoses in close proximity to surgery
Combinations and interactions	



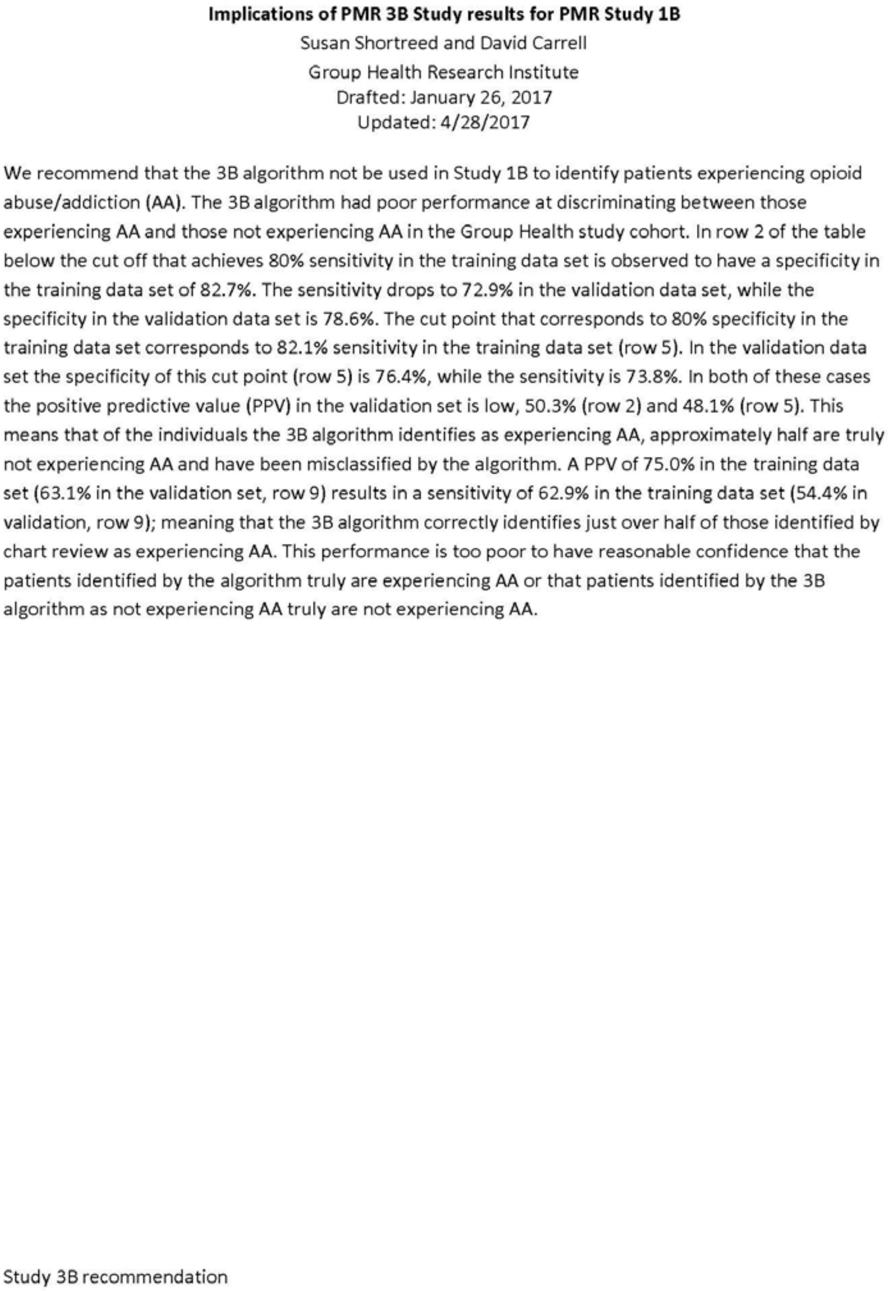
Combinations of data from multiple sources	Various versions, including frequency of urine drug screening during periods of overlapping opioid dispensings, emergency room encounters during periods of overlapping opioid dispensings, emergency room encounters during periods of excess days supply of opioids, emergency room encounters during concomitant use of opioids and other psychoactive medications, emergency room encounters during periods of high morphine equivalence dose
Interactions	Including interactions with patient age and interactions with patient gender

* For additional details see Appendix 3, "Potential Predictor Variables Considered for the 36-Month Claims-Based AA Model."

** Most potential predictors were derived in a variety of ways in both continuous and binary forms, including but not limited to: ever/never, frequency (overall, by month, by quarter), percent of time or visits, and/or in combination with other variables.

Source: Final report, pages 47-49.

7.2 APPENDIX B: INVESTIGATOR SUMMARY AND RECOMMENDATIONS.

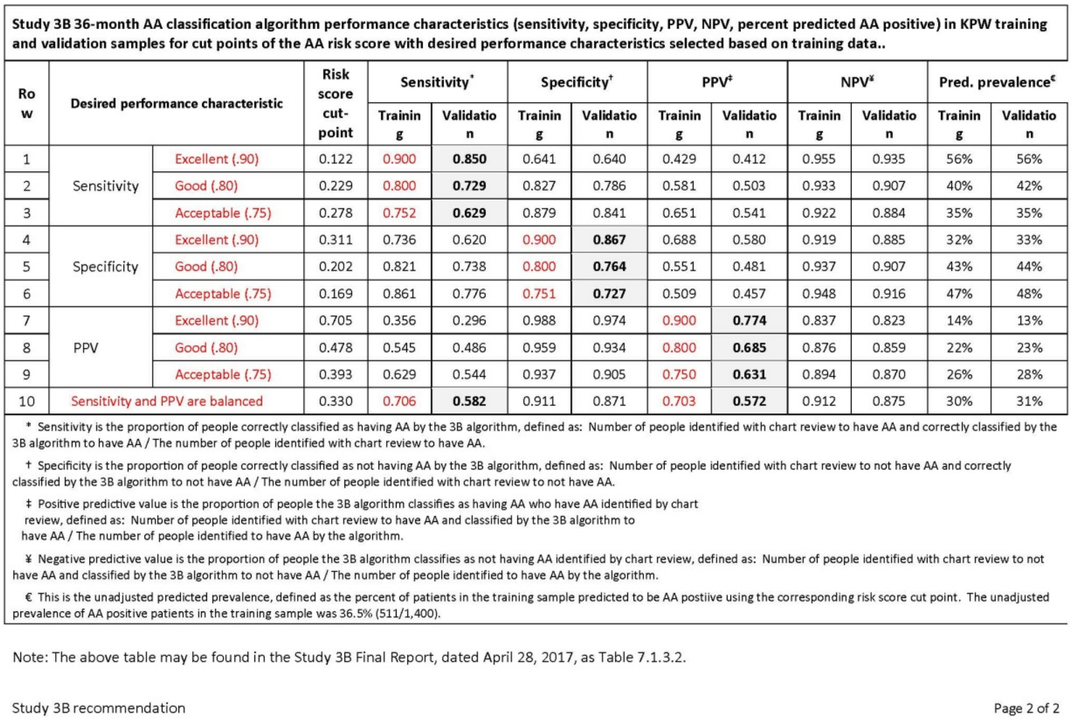


Implications of PMR 3B Study results for PMR Study 1B

Susan Shortreed and David Carrell
Group Health Research Institute
Drafted: January 26, 2017
Updated: 4/28/2017

We recommend that the 3B algorithm not be used in Study 1B to identify patients experiencing opioid abuse/addiction (AA). The 3B algorithm had poor performance at discriminating between those experiencing AA and those not experiencing AA in the Group Health study cohort. In row 2 of the table below the cut off that achieves 80% sensitivity in the training data set is observed to have a specificity in the training data set of 82.7%. The sensitivity drops to 72.9% in the validation data set, while the specificity in the validation data set is 78.6%. The cut point that corresponds to 80% specificity in the training data set corresponds to 82.1% sensitivity in the training data set (row 5). In the validation data set the specificity of this cut point (row 5) is 76.4%, while the sensitivity is 73.8%. In both of these cases the positive predictive value (PPV) in the validation set is low, 50.3% (row 2) and 48.1% (row 5). This means that of the individuals the 3B algorithm identifies as experiencing AA, approximately half are truly not experiencing AA and have been misclassified by the algorithm. A PPV of 75.0% in the training data set (63.1% in the validation set, row 9) results in a sensitivity of 62.9% in the training data set (54.4% in validation, row 9); meaning that the 3B algorithm correctly identifies just over half of those identified by chart review as experiencing AA. This performance is too poor to have reasonable confidence that the patients identified by the algorithm truly are experiencing AA or that patients identified by the 3B algorithm as not experiencing AA truly are not experiencing AA.

Study 3B recommendation



Study 3B 36-month AA classification algorithm performance characteristics (sensitivity, specificity, PPV, NPV, percent predicted AA positive) in KPW training and validation samples for cut points of the AA risk score with desired performance characteristics selected based on training data..

Row	Desired performance characteristic		Risk score cut-point	Sensitivity*		Specificity†		PPV‡		NPV¥		Pred. prevalence€	
				Training	Validation	Training	Validation	Training	Validation	Training	Validation	Training	Validation
1	Sensitivity	Excellent (.90)	0.122	0.900	**0.850**	0.641	0.640	0.429	0.412	0.955	0.935	56%	56%
2		Good (.80)	0.229	0.800	**0.729**	0.827	0.786	0.581	0.503	0.933	0.907	40%	42%
3		Acceptable (.75)	0.278	0.752	**0.629**	0.879	0.841	0.651	0.541	0.922	0.884	35%	35%
4	Specificity	Excellent (.90)	0.311	0.736	0.620	0.900	**0.867**	0.688	0.580	0.919	0.885	32%	33%
5		Good (.80)	0.202	0.821	0.738	0.800	**0.764**	0.551	0.481	0.937	0.907	43%	44%
6		Acceptable (.75)	0.169	0.861	0.776	0.751	**0.727**	0.509	0.457	0.948	0.916	47%	48%
7	PPV	Excellent (.90)	0.705	0.356	0.296	0.988	0.974	0.900	**0.774**	0.837	0.823	14%	13%
8		Good (.80)	0.478	0.545	0.486	0.959	0.934	0.800	**0.685**	0.876	0.859	22%	23%
9		Acceptable (.75)	0.393	0.629	0.544	0.937	0.905	0.750	**0.631**	0.894	0.870	26%	28%
10	Sensitivity and PPV are balanced		0.330	0.706	**0.582**	0.911	0.871	0.703	**0.572**	0.912	0.875	30%	31%

* Sensitivity is the proportion of people correctly classified as having AA by the 3B algorithm, defined as: Number of people identified with chart review to have AA and correctly classified by the 3B algorithm to have AA / The number of people identified with chart review to have AA.

† Specificity is the proportion of people correctly classified as not having AA by the 3B algorithm, defined as: Number of people identified with chart review to not have AA and correctly classified by the 3B algorithm to not have AA / The number of people identified with chart review to not have AA.

‡ Positive predictive value is the proportion of people the 3B algorithm classifies as having AA who have AA identified by chart review, defined as: Number of people identified with chart review to have AA and classified by the 3B algorithm to have AA / The number of people identified to have AA by the algorithm.

¥ Negative predictive value is the proportion of people the 3B algorithm classifies as not having AA identified by chart review, defined as: Number of people identified with chart review to not have AA and classified by the 3B algorithm to not have AA / The number of people identified to have AA by the algorithm.

€ This is the unadjusted predicted prevalence, defined as the percent of patients in the training sample predicted to be AA postiive using the corresponding risk score cut point. The unadjusted prevalence of AA positive patients in the training sample was 36.5% (511/1,400).

Note: The above table may be found in the Study 3B Final Report, dated April 28, 2017, as Table 7.1.3.2.

This is a representation of an electronic record that was signed electronically. Following this are manifestations of any and all electronic signatures for this electronic record.

/s/

CORWIN D HOWARD
04/29/2020 12:13:33 PM

CYNTHIA J KORNEGAY
05/05/2020 01:57:24 PM

JUDY A STAFFA
05/05/2020 02:36:12 PM